UNIVERSIDADE CATÓLICA PORTUGUESA

Potencialidades do *Power BI Desktop* na Análise Preditiva

por

Sofia Alexandra Santos Pinheiro

Católica Porto Business School

Agosto de 2020

UNIVERSIDADE CATÓLICA PORTUGUESA

Potencialidades do *Power BI Desktop* na Análise Preditiva

Trabalho Final na modalidade de Dissertação apresentado à Universidade Católica Portuguesa para obtenção do grau de mestre em Gestão na especialidade de Business Analytics

por

Sofia Alexandra Santos Pinheiro

sob orientação de

Professor Doutor António Andrade

Católica Porto Business School
Agosto de 2020

Resumo

Atualmente, é cada vez mais evidente e visível a quantidade de dados, que crescem a um ritmo exponencial dentro das organizações e, como tal, existe a necessidade de implementação de ferramentas de *Business Intelligence* (BI), capazes de fazerem o seu tratamento e posterior extração de informação e conhecimento, que permitam facilitar e dar suporte ao processo de tomada de decisão.

De forma complementar, surge o *Business Analytics*, que, aliado ao BI se traduz numa reforçada capacidade de análise e criação de valor para o negócio, que neste caso em concreto, se irá refletir na exploração da componente *Predictive Analytics* baseada em modelos de séries temporais.

O objetivo desta investigação foi o de aferir e identificar o potencial que as funcionalidades da ferramenta *Power BI Desktop*, na sua versão mais elementar, contribuem para um contexto de análise preditiva direcionada para utilizadores com conhecimentos superficiais de matérias de *Data Science*.

Foi, assim, concebido um tutorial explicativo e formativo das etapas imprescindíveis à realização de uma análise preditiva associado a um estudo de investigação exploratória, com testes aos diversos parâmetros de configuração, influenciadores da qualidade da mesma. Para este propósito, adotou-se o método *Knowledge Discovery in Databases* (KDD) no que respeita ao tratamento de dados e partiu-se do modelo *Selecting, Organizing, Integrating* (SOI) como base para a elaboração do tutorial.

A análise realizada permitiu verificar que existem três parâmetros essenciais, com influência direta no desempenho do algoritmo de previsão e consequente qualidade da mesma, sendo estes o ignorar último, o intervalo de confiança e a sazonalidade.

Palavras-Chave: Business Analytics, Business Intelligence, Power BI, Análise Preditiva

Abstract

Nowadays, it is increasingly evident and visible the amount of data, which grows at an exponential pace within organizations and, as such, there is a need to implement Business Intelligence (BI) tools, capable of making their treatment and later extraction of information and knowledge, to facilitate and support the decision-making process.

In a complementary way, Business Analytics emerges, which, combined with BI, translates into an enhanced capacity for analysis and value creation for the business, that in this specific case, will be reflected in the exploration of the Predictive Analytics component based on time series models.

The purpose of this investigation was to assess and identify the potential of Power BI Desktop's functionalities, in its most elementary version, contribute to a context of predictive analysis aimed at users with superficial knowledge of Data Science subjects.

Thus, an explanatory and formative tutorial of the essential steps to carry out a predictive analysis associated with an exploratory research study was designed, with tests on the various configuration parameters, influencing the quality of the prediction. For this purpose, the Knowledge Discovery in Databases (KDD) method was implemented regarding data treatment and the selecting, organizing, integrating (SOI) model was used as the basis for the elaboration of the tutorial.

The analysis made it possible to verify that there are three essential parameters, with direct influence on the performance of the forecasting algorithm and its consequent quality, such the ignore last, the confidence interval and seasonality.

Keywords: *Business Analytics, Business Intelligence, Power BI, Predictive Analytics*

Índice

Resumo v
Abstract vii
Índice ix
Índice de figuras xiii
Índice de quadros xvi
Lista de siglas, abreviaturas e acrónimos xviii

Introdução 1
Enquadramento e Motivação 1
Questão de Investigação 3
Método de Investigação 4
Estrutura da Dissertação 4

Business Intelligence 7
1. Evolução, conceitos e características 7
1.1 Dos dados ao conhecimento 7
1.2 Evolução do conceito de BI 9
1.3 Arquitetura de um sistema de BI 12
1.4 Ferramentas de BI líderes de mercado 18
1.5 Benefícios dos sistemas de BI nas organizações 20
1.6 Business Analytics 22
1.7 Análise preditiva 24
1.7.1 Conceito 24
1.7.2 Métodos e tipologia de dados 25
1.7.3 Métricas de erro 30

Microsoft Power BI 32
2. O recurso tecnológico Power BI 32
2.1 Caracterização global 32
2.2 Componente Desktop 35
2.3 Utilizadores Power BI 36
2.4 Potencialidades 38
2.5 Recursos complementares ao Power BI 41
2.5.1 Integração de contributos de parceiros da Microsoft 41
2.5.2 Integração de contributos da Microsoft 42

Design Metodológico 45
3. Métodos de análise de dados e conceção de LO 45
3.1 Métodos de análise de dados 45
3.1.1 CRISP-DM 46
3.1.2 SEMMA 48
3.1.3 KDD 50
3.2 Conceção de LO 53
3.3 Design metodológico adotado 55

Ensaio do Power BI na análise preditiva 56
4. Exploração do Power BI Desktop 56
4.1 Enquadramento 56
4.2 Análise preditiva com recurso ao line chart 60
4.3 Aplicação prática em PBI Desktop 65
4.3.1 Tutorial 65
4.3.2 Interpretação e avaliação 72
4.3.2.1 Teste ao parâmetro ignorar último 73
4.3.2.2 Teste ao parâmetro intervalo de confiança 79
4.3.2.3 Teste ao parâmetro sazonalidade 84

Conclusão 88
5. Síntese da investigação, limitações e trabalho futuro 88
5.1 Síntese da investigação 88
5.2 Trabalho futuro 91

Referências Bibliográficas 92
Apêndice A - Business Intelligence 100
A.1 Evolução do conceito de BI 100
A.2 ETL - Definições, objetivos e características 101
A.3 Data Warehouse e Data Mart 103
A.4 Servidor On-line Analytical Processing (OLAP) 108
A.5 Análise preditiva 110
A.5.1 Simple Exponential Smoothing 110
A.5.2 Double Exponential Smoothing 110
A.5.3 Triple Exponential Smoothing 111

Apêndice B - Ensaio do Power BI na análise preditiva 113
Apêndice C - Organização do site de introdução ao Power BI 118

C.1 Organização global 118
C.2 Vídeos 119
C.2.1 Vídeo do tutorial 120
C.2.2 Vídeo do teste ao parâmetro ignorar último 121
C.2.3 Vídeo do teste ao parâmetro intervalo de confiança 121
C.2.4 Vídeo do teste ao parâmetro sazonalidade 122

Índice de figuras

Figura 1- Dados, Informação e conhecimento 9
Figura 2 - Arquitetura de um sistema-padrão de BI..... 13
Figura 3 – Quadrante Mágico de Gartner, 2019 19
Figura 4 - Tipos de utilizadores de PBI 37
Figura 5 - Etapas do ciclo CRISP-DM 48
Figura 6 - Etapas constituintes do processo SEMMA..... 50
Figura 7- Estágios do processo KDD 53
Figura 8- Conexão à fonte de dados..... 66
Figura 9-Confirmação do formato das variáveis 67
Figura 10- Line chart ou gráfico de linhas..... 67
Figura 11- Hierarquia de calendário 68
Figura 12- Opção expandir tudo..... 68
Figura 13- Lupa de análise 69
Figura 14- Linha de previsão..... 70
Figura 15 - Parâmetros de configuração..... 71
Figura 16- Exportar dados 72
Figura 17- Modelo 1 do teste ao parâmetro IU..... 74
Figura 18- Modelo 2 do teste do parâmetro IU 75
Figura 19 - Modelo 3 do teste ao parâmetro IU..... 76
Figura 20 – Modelo 4 do teste ao parâmetro IU 77
Figura 21 - Modelo 1 do teste ao parâmetro IC 79
Figura 22 - Modelo 2 do teste ao parâmetro IC 80
Figura 23 - Modelo 3 do teste ao parâmetro IC 80
Figura 24 - Modelo 4 do teste ao parâmetro IC 81
Figura 25 - Modelo 5 do teste ao parâmetro IC 83

Figura 26 - Modelo 6 do teste ao parâmetro IC 83
Figura 27 - Modelo 1 do teste ao parâmetro sazonalidade 84
Figura 28 - Modelo 2 do teste ao parâmetro sazonalidade 85
Figura 29 - Modelo 3 do teste ao parâmetro sazonalidade 85
Figura 30 - Modelo 4 do teste ao parâmetro sazonalidade 86
Figura 31 – Layout geral do site 118
Figura 32 – Separador Predictive Analytics 119
Figura 33 – Separador Tutorial 119
Figura 34 – Separador Vídeos 120
Figura 35 – Vídeo demonstrativo do tutorial 120
Figura 36 - Vídeo do teste ao parâmetro ignorar último 121
Figura 37 – Vídeo do teste ao parâmetro intervalo de confiança 121
Figura 38 – Vídeo do teste ao parâmetro sazonalidade 122

Índice de quadros

Quadro 1- Métricas de Erro 31

Quadro 2 – Quadro de erros .....

Quadro 3 – Upper bound e lower bound do IC .....

Quadro 4 – Conclusão por objetivo proposto 91

Quadro 5 – Comparação de funcionalidades entre versões do Power BI 117

Lista de siglas, abreviaturas e acrónimos

AML – Azure Machine Learning

BA – Business Analytics

BD – Base de Dados

BI – Business Intelligence

CRISP-DM – Cross-Industry Standard Process for Data Mining

DSA - Data Staging Area

DAX – Data Analysis eXpression

DM – Data Mining

DW – Data Warehouse

ESM – Exponential Smoothing Model

ETL – Extraction, Transformation, Loading

HT – Horizonte Temporal

IA – Inteligência Artificial

IC – Intervalo de Confiança

IOT – Internet of Things

IU – Ignorar Último

KDD – Knowledge Discovery in Databases

KPI – Key Performance Indicator

LO – Learning Object

ML – Machine Learning

MS – Microsoft

OLAP – On-line Analytical Processing

PBI – Power BI

SEMMA – Sample, Explore, Modify, Model, Assess

SGBD – Sistema de Gestão de Base de Dados

SI – Sistemas de Informação

SOI – *Selecting, Organizing, Integrating* ou Seleção, Organização, Integração

SQL – Structured Query Language

SSBI – Self-Service BI

SSD – Sistemas de Suporte à Decisão

TI – Tecnologias de Informação

Introdução

Enquadramento e Motivação

A progressiva digitalização da vida empresarial e pessoal que se acelera com a integração de diversos subsistemas de informação, que permite a interação das pessoas com os mesmos e a progressiva ligação de todas as coisas (Internet of Things) gera o designado *Big Data*. Desta forma, empresas pequenas, médias ou de grande dimensão acabam por ver nestes dados, por um lado a possibilidade de melhorar o controlo e gestão da sua atividade e, por outro, apoiar o seu sistema de suporte à decisão (Antonelli, 2009).

A necessidade de analisar e cruzar dados internos estruturados e transacionais (produção, compras, vendas, recursos humanos, etc.) com dados não estruturados internos (ex. email) e externos (ex. plataformas sociais) tem emergido na atualidade.

Paralelamente, tem-se testemunhado um crescimento e evolução a nível de ferramentas e sistemas de apoio à decisão, surgindo os designados sistemas de *Business Intelligence*, provenientes sobretudo de pressões e fatores externos nomeadamente por parte de *stakeholders*[1] e da necessidade de diferenciação e antecipação das empresas face às concorrentes.

Contudo, esta pressão revelou-se pertinente e urgente em encontrar soluções em tempo real e viáveis que permita à organizações tomar decisões de longo prazo de forma otimizada e consciente, conferindo-lhes segurança, fiabilidade, qualidade e precisão no que diz respeito a análise de tendências, às transformações que ocorrem no mercado, padrões de consumo, preferências de clientes, entre outros indicadores (Sezões *et al*., 2006), (Olszak & Ziemba, 2007).

[1] *Stakeholders* são grupos de intervenientes com interesse nos processos e resultados de dada organização

Claramente que, num cenário atual representativo da era tecnológica aliada à transformação digital, revela-se imperativo e crucialmente relevante que todos os tipos de negócio comecem a adotar de forma correta, adaptada e assertiva ao seu contexto organizacional algumas ferramentas de *Business Intelligence* no mercado que facilitem os seus propósitos.

Na prática, os sistemas de *BI* supramencionados recolhem os dados, armazenam-nos em repositórios (que será pormenorizadamente explicado no subcapítulo referente à arquitetura), onde são analisados. Posteriormente, os mesmos são submetidos a uma panóplia de algoritmos sofisticados através de diversas ferramentas de *BI*, que auxiliam na análise dos dados e, que, numa fase mais madura permitem que se extraia a informação necessária que por sua vez será transformada em conhecimento. Conhecimento este extremamente importante e imprescindível que suporta a tomada de decisões que contribuirá para o alcance dos objetivos pretendidos da organização e do seu negócio, conseguindo desta forma, acompanhar e colmatar as necessidades dos seus clientes (Negash & Gray, 2003), (Santos & Ramos, 2017).

Assim, é necessário que exista um alinhamento entre a tecnologia e a própria estratégia das organizações com vista a melhorar a gestão dos agentes responsáveis, criando valor para o panorama geral, inovando e diferenciando-se das demais empresas existentes (Sezões *et al.*, 2006), (Vercellis, 2009).

É esta simbiose que leva a uma melhor execução por parte das empresas, que por sua vez leva à criação de valor e consequente vantagem competitiva, já que estes sistemas de *BI* representam o ponto de viragem, abrindo caminho para a adesão a novas práticas de gestão que acaba por se refletir em decisões melhor fundamentadas, melhores resultados e melhor desempenho (Sezões *et al.*, 2006).

Questão de Investigação

O objetivo central deste estudo prende-se com a identificação e a compreensão do potencial e aptidões da ferramenta *Power BI* que se encontram ao alcance de utilizadores sem conhecimentos rigorosos de *SQL*, estatística e matemática, num contexto de análise preditiva, materializando-se na seguinte questão:

"Qual o potencial do *Power BI* para a análise preditiva"?

Estão fora deste cenário exploratório enquadramentos empresariais mais sofisticados com acesso ao *Power BI Premium*, a *SQL* com os seus cubos *OLAP*, e demais recursos especificamente desenhados para potenciar o *Business Intelligence* e aferir as suas potencialidades preditivas.

Para definir o caminho de pesquisa formularam-se um conjunto de objetivos específicos que se explicitam:

- Identificar recursos e potencialidades disponíveis no *Power BI* para o efeito;
- Pesquisar recursos adicionais integráveis no *Power BI* para análise preditiva;
- Identificar competências essenciais para aplicar em contexto de *Power BI* na análise preditiva;
- Explorar experimentalmente os recursos, de forma a compreender, na prática, o impacto e contributo dos mesmos neste tipo de análise;
- Avaliar o rigor, qualidade e suporte que a ferramenta poderá conferir aos seus utilizadores no âmbito da previsão;
- Detetar benefícios, fragilidades e desafios no âmbito da previsão e do *Power BI.*

A explicitação do potencial identificado no contexto técnico definido de alguma simplicidade e numa abordagem para utilizadores genéricos, não especialistas, será materializado pela construção de objetos formativos (LO) que

facilitem a compreensão da ferramenta e acelere a sua eventual adoção por empresas de menor dimensão e com recursos técnicos e humanos mais limitados.

Método de Investigação

O processo de investigação recorrerá aos princípios do *Design Science Research*, onde se escolherá e aplicará o método de tratamento de dados e de conceção de *LO* mais adequados. Numa fase posterior, estes métodos serão devidamente explicados, bem como os devidos algoritmos apropriados que o *Power BI* disponibiliza para aplicação a fontes de dados simplificadas. O objetivo, recorde-se, é o de explicar e identificar de uma forma demonstrativa as diversas potencialidades da ferramenta, sob a forma de *LO* expostos em pequenos tutoriais num *site* especificamente criado para servir este propósito.

Contudo, esta abordagem será desenvolvida e aprofundada no capítulo respeitante ao *Design* Metodológico.

Estrutura da Dissertação

Para se atingir o objetivo desta investigação, é necessário efetuar uma revisão sobre a literatura referente aos vários conceitos ligados a este tema, concretizando-se sob a forma de um estudo mais profundo acerca dos conceitos *Business Intelligence* e *Power BI*. Para tal, recorrer-se-á à compreensão de alguns conteúdos de *Business Analytics* e *Data Mining*, nomeadamente no que a métodos e algoritmos de previsão diz respeito e, paralelamente, identificar potenciais vantagens e fragilidades para os utilizadores.

No que respeita à organização deste estudo, segue-se uma breve explicação da sua estrutura:

O capítulo 1 constitui a revisão de literatura, onde será feita uma introdução ao *Business Intelligence*, detalhando o seu conceito e respetiva evolução, as suas características principais, a sua arquitetura, os diversos utilizadores, as vantagens da sua adoção, bem como uma comparação superficial entre as ferramentas de *BI* líderes de mercado. Adicionalmente, será feita uma alusão ao *Business Analytics* e à sua componente de análise preditiva.

No capítulo 2 será efetuado um estudo pormenorizado a uma das ferramentas de *BI* líderes de mercado, *Power BI*, onde se irá explorar a sua caracterização global, a sua componente *Desktop*, bem como os principais utilizadores, as suas funcionalidades e recursos complementares considerados relevantes.

No capítulo 3, expor-se-á de uma forma pragmática os métodos de tratamento de dados existentes, bem como o método de conceção dos LO, onde, posteriormente, se irá escolher e detalhar aquele que melhor se adequa ao propósito.

O capítulo 4 debruçar-se-á na exploração prática da ferramenta PBI. Inicialmente, será detalhada a aplicação do método de tratamento de dados bem como da conceção de LO implementados. De seguida, explicar-se-á, de forma teórica, o funcionamento da ferramenta na análise em questão, concentrando estes conceitos na forma de um tutorial representativo das fases fundamentais à sua concretização. Adicionalmente, serão gerados alguns modelos de teste aos diversos parâmetros, com o fim de avaliar o seu contributo e relevância.

O capítulo 5 constitui a conclusão deste estudo, que será acompanhada por uma síntese do trabalho elaborado, uma análise aos objetivos pretendidos, das metodologias postas em prática, bem como dos resultados atingidos. Será feito, ainda, um levantamento das limitações e sugestões relativamente a trabalho futuro que possa ser efetuado.

Por fim, no seguimento, encontrar-se-á uma descrição aprofundada de alguns dos tópicos identificados nos capítulos 1, 2 e 3, na forma de Apêndices, com o objetivo de compreender o objeto de estudo.

Capítulo 1
Business Intelligence

1. Evolução, conceitos relacionados e características

1.1 Dos dados ao conhecimento

Antes de se partir para um estudo profundo acerca dos sistemas de *Business Intelligence* e do próprio *Business Intelligence*, é importante perceber e ter uma noção de alguns conceitos e termos utilizados e relacionados com o *BI*. Importa, por isso, compreender de forma breve o que são na realidade os **dados** e sistematizar as principais diferenças entre a **informação** e **conhecimento**.

Relativamente aos **dados**, por definição, podem ser entendidos como eventos, factos e itens elementares, que de forma isolada não apresentam qualquer significado (Goldschmidt & Passos, 2005). É necessário que exista uma quantidade considerável de dados que consigam ser objeto de estudo (Antonelli, 2009).

A **informação** e passando a citar Le Moigne é um:

"objeto formatado, criado artificialmente pelo homem, tendo por finalidade representar um tipo de acontecimento identificável por ele no mundo real, integrando um conjunto de registos ou dados e um conjunto de relações entre eles, que determinam o seu formato" (Moigne, 1978).

Marta Valentim descreve a informação como sendo dados relacionados e processados de forma a que sejam significativos (Valentim, 2002). Com isto, depreende-se que a informação consiste no *output* gerado, a partir da análise e extração efetuada ao conjunto de dados recolhidos e organizados, que são

posteriormente transformados em informação útil, relevante e aplicável a um cenário ou contexto específico com significado para o negócio (Antonelli, 2009), (Sezões *et al*., 2006), (Vercellis, 2009), (Goldschmidt & Passos, 2005).

O papel principal da informação, tal como o seu nome indica é informar pessoas e processos, fornecendo factos e indicadores fulcrais tanto para os processos de negócio como para os agentes de negócio (Sezões *et al*., 2006).

O **conhecimento**, por sua vez, consiste numa combinação de ideias, regras e procedimentos que constituem a base e auxiliam no desenvolvimento de ações e tomada de decisões (Antonelli, 2009). Para tal, é necessário que a informação seja transformada em conhecimento ou que o conhecimento consiga ser extraído dos dados armazenados. Assim, estipulam-se duas formas principais de o obter: a forma passiva, onde se enfatizam os critérios de análise seguidos pelos agentes decisivos, e, em alternativa, é sugerido que se utilizem determinadas ferramentas e modelos de análise de dados que, de forma ativa, são aplicados aos mesmos para esse fim. Claramente que toda esta tecnologia inerente à integração de processos e gestão de conhecimento é aliada à experiência e competência dos principais responsáveis pela tomada dessas decisões, que acabam por lhes facilitar o processo de escolha, ao conferirem um maior apoio, confiança e certeza (Vercellis, 2009). A **Figura 1** ilustra os conceitos explicados.

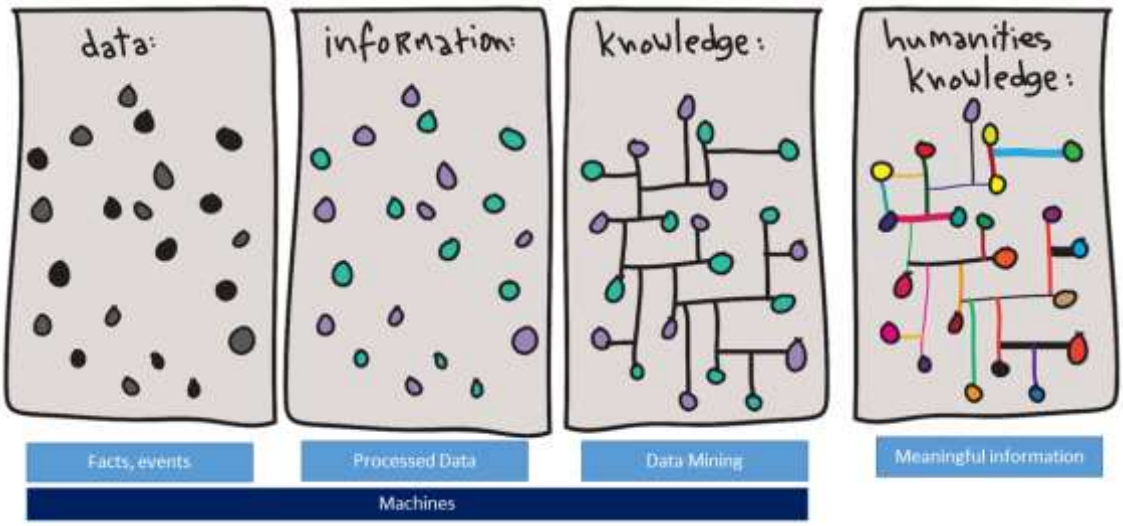

Figura 1- Dados, Informação e conhecimento

Fonte: Deb Verhoeven, 2015[2]

1.2 Evolução do conceito de *Business Intelligence*

Foi essencialmente na década de 1980 que surgem pela primeira vez os sistemas de *Business Intelligence*, devido sobretudo à evolução dos Sistemas de Suporte à Decisão (SSD) e progresso tecnológico ao nível dos computadores pessoais e do aumento da capacidade de processamento dos mesmos. Este tipo de sistemas objetiva controlar e apoiar nas tomadas de decisão, combinando, para tal, o *Data Warehouse*, as plataformas e ferramentas de *BI*, com o fim de conseguirem transformar os dados em informação útil para o negócio. Desta forma, evidencia-se a existência de dois subsistemas fundamentais nas organizações: o controlo e a tomada de decisão (Zorrinho, 1991).

Paralelamente, tem-se testemunhado um elevado crescimento e evolução a nível deste tipo de ferramentas e sistemas de apoio à decisão, já que constituem novas formas de armazenamento de grandes volumes de dados e são

[2] Consultado em https://figshare.com/articles/data_information_knowledge_humanities/1397453 a 10/12/2019

ferramentas mais simples, eficazes e expeditas de extração, limpeza e análise de dados (Elena, 2011), (Vercellis, 2009), (Negash & Gray, 2003).

A par dos sistemas de *BI* surge o conceito de *Business Intelligence*, que, de uma forma muito breve, constitui todo o conjunto de tecnologias, instrumentos e aplicações que, auxiliando a análise e exploração de grande volume de dados, os transforma em informação. Por sua vez, esta, é convertida em conhecimento útil, acabando por se refletir em decisões devidamente fundamentadas, otimizando o desempenho das organizações (Sezões *et al*., 2006), (Antonelli, 2009), (Borges, Cardozo & Filho, 2018), (Soares & Silva, s.d.).

O conceito de *Business Intelligence* está em constante evolução e tem ganho reconhecimento mundial ao longo do tempo. Tal, acaba por se refletir num vasto número de definições propostas e apresentadas por diversos autores (Antonelli, 2009).

Hans Peter Luhn, investigador da *IBM* é considerado o autor pioneiro do conceito de *BI* ao elaborar o artigo "A Business Intelligence System" em 1958. Luhn atribuía especial atenção à comunicação dentro de uma organização, mais concretamente à expansão e disseminação da informação por toda a organização, já que, para Luhn, a "comunicação eficiente é uma chave para o progresso em todos os campos do esforço humano" (Luhn, 1958). À luz do conceito do mesmo, este via a comunicação como um meio facilitador para a condução de um dado negócio dentro das organizações, tentando ao máximo melhorá-la e tirar o maior proveito dela. Desenvolveu assim, um sistema automático, baseado em máquinas de processamento de dados responsável por fazer a triagem de documentos, expandir e difundir informações por toda a organização, com o objetivo de conseguir obter informações úteis sobre o negócio da mesma (Luhn, 1958), (Elena, 2011).

Em 1989 surge a segunda definição de *BI* concebida por **Howard Dresner**, um investigador integrante do *Gartner Group.* Segundo a sua visão, o *BI* é comparado com a ideia de um guarda-chuva de conceitos, métodos e informações a serem

tidos em conta para aperfeiçoar e agilizar o processo de tomada de decisões empresariais, recorrendo a sistemas de suporte baseados em factos (Elena, 2011).

Também **Barbieri** segue a mesma linha de pensamento de Dresner e define o *BI* como um guarda-chuva conceitual, enfatizando a fase de recolha de dados, informações e conhecimento que permitam às organizações maximizarem a sua competitividade (Barbieri, 2001). Este autor dá especial enfoque aos repositórios de dados, nomeadamente ao *Datawarehouse*, *Datamarts*, bem como às técnicas de análise dos mesmos, como forma de otimizar a *performance* e a tomada de decisões em tempo real (Barbieri, 2001). Para concluir a sua premissa, Barbieri (2001) defende que o *BI* se dedica à utilização de diversas fontes de informação, culminando em estratégias de competitividade para os negócios da empresa, utilizando para tal fim, métodos que procuram encontrar relações, padrões, correlações ou tendências camuflados e implícitos nesses dados.

O ***Gartner Group*** [3]foi também considerado um grupo muito mediático que gerou um grande contributo ao personificar o que muitos designam por "pai" do *BI*, tendo tido um enorme peso na forma como os sistemas de *BI* são vistos atualmente e na sua difusão. Ainda neste seguimento, Gartner concorda com o termo "guarda-chuva" no que respeita à sua definição, defendendo que consiste na aplicação de um conjunto de metodologias e tecnologias como ferramentas de análise, OLAP, *data warehouse*, *data mining*, *predictive analytics*, entre outras práticas implementadas e desenvolvidas numa organização que conduzem à otimização de decisões e consequentemente da sua vantagem competitiva (Ranjan, 2005).

De um modo geral, todos os conceitos e definições expostos e formulados pelos mais diversos autores apresentam-se de forma extremamente alinhada, partilhando vários objetivos e ideias em comum, podendo outras definições serem consultadas no Apêndice A.

[3] O Gartner Group é uma empresa de consultoria fundada em 1979 por Gideon Gartner que se dedica ao desenvolvimento de tecnologias que dão suporte aos processos de tomada de decisões

Para concluir, os sistemas de *BI* são então responsáveis por todo o processo de recolha de dados, armazenamento, análise e exploração dos mesmos que, através de tecnologia e metodologia adequadas, têm em vista obter *outputs*. Estes, refletem-se em conhecimento eficiente, útil, de boa qualidade e contextualizado, facilitando a interpretação dos dados e a gestão de conhecimento dentro das organizações, que por sua vez dispõem de inúmeras ferramentas de análise que potenciam e melhoram o seu processo de decisão e consequente implementação de estratégias que levam a adquirirem vantagem competitiva (Negash & Gray, 2003), (Watson & Wixom 2007).

O objetivo central das soluções de *BI* é o de facilitar a compreensão do negócio das organizações, fornecendo informações relevantes e cruciais a todos os níveis da organização, nomeadamente sobre as suas operações internas e ambiente externo, incluindo os seus clientes, concorrentes, parceiros e fornecedores, constituindo assim, uma mais valia para qualquer organização e para qualquer processo de tomada de decisão bem fundamentado (T.Moss & Atre, 2003), (Sezões *et al.*, 2006).

1.3 Arquitetura de um sistema de BI

Antes de se implementar qualquer tipo de sistema de *BI* numa dada organização, é de grande importância identificar as necessidades reais da empresa, a estratégia adotada, bem como os seus objetivos a atingir, de forma a que se consiga encontrar um sistema que providencie ferramentas e soluções que estejam alinhadas com esse fim.

Um sistema de *BI* deve estar oportunamente enquadrado e relacionado com a infraestrutura global dos sistemas de uma organização e se, por um lado não se pode dissociar das fontes de dados inerentes, por outro, deve atribuir-se especial

importância se os *outputs* gerados para os seus utilizadores estão em conformidade com o requerido por estes (Sezões *et al*., 2006).

A arquitetura de um sistema-padrão de *Business Intelligence* proposto por Han, Kamber e Pei deve contemplar os elementos que estão exemplificados na **Figura 2**. Estas componentes constituem ferramentas que, operando em conjunto, são responsáveis pelo armazenamento dos dados, análise de informações e mineração dos dados (Antonelli, 2009), (Sezões *et al*., 2006), (Han *et al*., 2012).

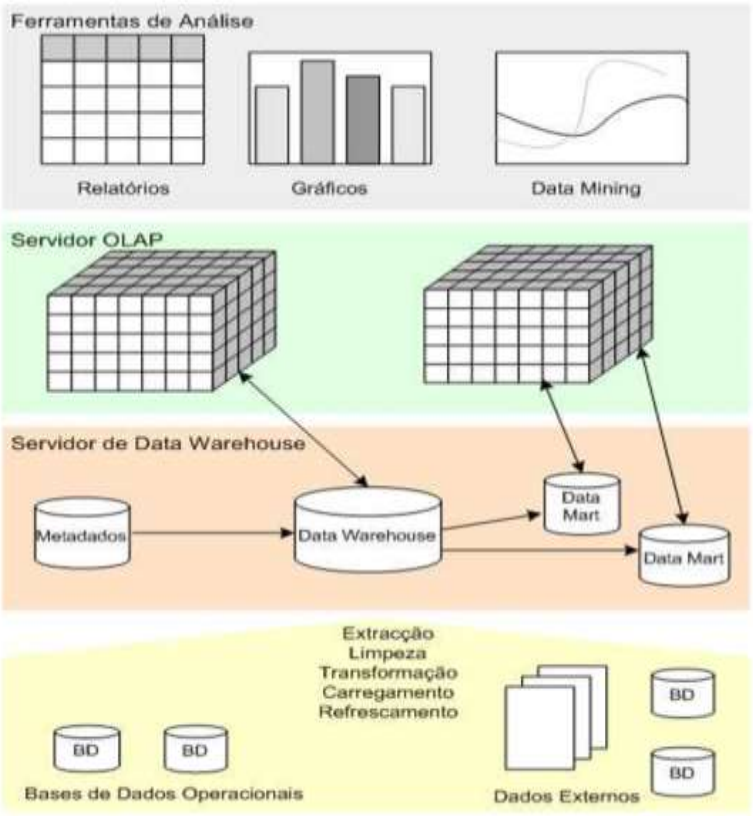

Figura 2 - Arquitetura de um sistema-padrão de BI

Fonte: Han, Kamber & Pei, 2012

Como se pode observar na **Figura 2**, os designados componentes dividem-se em quatro níveis principais, nomeadamente o nível inferior: servidor de *Data Warehouse* que contém as aplicações de suporte ao processo ETL (Extração,

Transformação e Carregamento), que na sua explicação detalhada vão ser dissociados, perfazendo os ditos quatro níveis. O nível intermédio constituído pelo servidor de *On-Line Analytical Processing* (OLAP) e por último, o nível superior que se refere às ferramentas de análise e à disponibilização de informação relevante resultante, sendo estes brevemente abaixo descritos: (Santos & Ramos, 2017), (Han & Kamber & Pei, 2011).

- **Componentes de *ETL*** (Extration, Transformation, Loading), comummente designadas por ferramentas de *Back End,* que tal como o nome indica, se dedica à extração, transformação e carregamento dos dados. É a componente responsável pela recolha de informações que sejam potencialmente necessárias para todo o processo de decisão, provenientes de diversas fontes de dados, podendo estas ser externas ou internas (sistemas ERP, arquivos TXT, ficheiros excel, entre outros) a serem posteriormente armazenados no *Data Warehouse* (Antonelli, 2009), (Sezões *et al*., 2006).
- **Servidor de *Data warehouse*/*Data marts*** – É um armazém de dados, um repositório integrado, onde ficam concentrados e armazenados todos os dados extraídos dos sistemas operacionais, através da integração de um Sistema de Gestão de Bases de Dados (SGBD) e de um conjunto de ferramentas de suporte ao processo *ETL*. Esses dados provenientes das Bases de Dados (BD) operacionais, e de outras fontes de dados internas ou externas à organização, são carregados para o *DW* através de API *(Application Program Interfaces),* após a sua passagem pelo processo de limpeza e de transformação, de acordo com as etapas constituintes de *ETL,* para que apresentem o formato dos dados do DW. Das aplicações mais utilizadas no processo de carregamento destacam-se as ligações ODBC *(Open Database Connectivity),* que

permitem o acesso aos dados nos sistemas-fonte e a sua posterior transferência (Santos & Ramos, 2017).

- **Servidor OLAP (*On-line Analytical Processing)*** – É uma ferramenta que permite efetuar uma análise multidimensional aos dados provenientes do *DW*, na medida em que são visualizados diversos cubos analíticos, permitindo desta forma, examinar a informação sob diferentes perspetivas. A implementação deste servidor pode ser feita através de três modelos: modelo ROLAP, MOLAP ou ainda HOLAP que estão devidamente detalhados no apêndice A (Santos & Ramos, 2017).
- **Nível de resultados/ferramentas** – O nível superior, também designado por *Front End* comporta todas as ferramentas de análise utilizadas para gerar relatórios, identificar tendências ou padrões nos dados. Refere-se, portanto, a todas as ferramentas de *Data Mining* ou melhor dizendo, a todos os métodos, processos e algoritmos de exploração de dados, com o objetivo principal de identificação de padrões, relacionamentos, tendências, modelos, projeções de cenários futuros, entre outros, que muitas das vezes não se evidenciam na conjuntura dos dados. A informação relevante que daí advém é providenciada, sob a forma de relatórios, gráficos, modelos, entre outros (Santos & Ramos, 2017).

Detalhando melhor este último nível, o *Data Mining* é um conceito de descoberta de conhecimento nos dados que vulgarmente se caracteriza como "mineração de dados". É assim um processo de exploração e análise de um grande volume de dados que, objetiva de forma clara e precisa, descobrir padrões, tendências e relacionamentos implícitos ou camuflados nesses dados. A partir daí, visa ainda extrair conhecimento, que se possa refletir em relações, modelos e associações, por forma a facilitar, sustentar e apoiar a tomada de

decisão nas empresas, recorrendo, para o efeito, à utilização de diferentes técnicas e algoritmos (Vercellis, 2009), (Primak, 2008), (Antonelli, 2009), (Sezões *et al*., 2006), (Camilo & Silva, 2009) (Singh *et al*., 2019), (Santos & Ramos, 2017), (Berry & Linoff, 2004) (Hand, 2007). É possível aferir a importância que o *Data Mining* apresenta, sendo uma ferramenta bastante reconhecida a nível dos resultados, registando consequentemente uma elevada taxa de adesão no mundo empresarial (Antonelli, 2009), (Sezões *et al*., 2006).

Contudo, é importante salientar que o *Data Mining* não substitui o fator humano no que respeita à análise de dados, ou seja, é necessário que os resultados provenientes deste processo coexistam com a capacidade de avaliação e análise dos gestores de forma a determinar o valor e o real impacto para o negócio (Sezões *et al*., 2006), (Camilo & Silva, 2009).

O *Data Mining* tem a si associadas determinadas tarefas de aprendizagem que se podem dividir em dois grupos: a **descrição**, que inclui tarefas de **aprendizagem não supervisionada**, e a **previsão** que, por sua vez, engloba **tarefas de aprendizagem supervisionada**. A aprendizagem não supervisionada, caracteriza-se por não existir qualquer conhecimento relativo aos dados e como tal, o utilizador apresenta uma atitude passiva, uma vez que os dados por si só definem as classes (Santos & Ramos, 2017), (Gama, Carvalho, Faceli, Lorena & Oliveira, 2015). Por outro lado, a aprendizagem supervisionada pressupõe que exista um conhecimento prévio dos dados, bem como dos atributos e classes, permitindo assim ao analista o auxílio na construção do modelo que se pretende.

No que concerne às tarefas de **descrição**, o objetivo é o de explorar, interpretar e caracterizar um dado conjunto de dados, de forma a identificar regras, associações ou relações entre eles, permitindo aumentar assim, o nível de conhecimento dos mesmos. Nesta categoria, encontram-se as tarefas de **segmentação (*clusters*)**, **associação** e **sumarização** (Santos & Ramos, 2017) (Camilo & Silva, 2009), (Gama *et al*., 2015), (Berry & Linoff, 2004).

No que respeita a tarefas de **previsão**, a sua finalidade é a de formular funções (modelos ou hipóteses) suficientemente aptos e qualificados na previsão de valores futuros de uma dada variável ou atributo *(*Gama *et al*., 2015*)* (Santos & Ramos, 2017), (Camilo & Silva, 2009), (Sezões *et al*., 2006), (Vercellis, 2009), (Berry & Linoff, 2004). Se a variável preditiva for uma variável numérica contínua, o modelo de previsão apropriado será o da **regressão**. Por outro lado, se a variável preditiva for uma variável categórica, discreta, o modelo de previsão adequado é a **classificação**, que aplica o mesmo princípio da regressão, alterando apenas a natureza da variável preditiva. Por último, se a variável preditiva for *time dependent* e por isso, apresentar um histórico temporal, então o método preditivo adequado é a previsão em **séries temporais** (Delen, 2015). Posteriormente, esses modelos são testados a nível da sua fiabilidade e validados quando comparados com a realidade, contribuindo para sustentar o processo de tomada de decisão (Sezões *et al*., 2006), (Santos & Ramos, 2017), (Vercellis, 2009). Ao contrário da descrição, a previsão é um processo de natureza cíclica que deverá estar em constante avaliação, acompanhando desta forma as constantes modificações das variáveis de negócio (Sezões *et al*., 2006).

Contudo, é de ressalvar que a divisão acima descrita entre modelos preditivos e descritivos nem sempre é assim tão definida, já que um modelo preditivo também poderá ter capacidade de descrição de um conjunto de dados e, por sua vez, um modelo descritivo após validação, também poderá efetuar previsões (Gama *et al*., 2015).

Por último, é vastíssimo o leque de técnicas ou métodos de *Data Mining* que podem ser aplicadas de acordo com o tipo de tarefa de aprendizagem, recorrendo para tal a diferentes algoritmos utilizados em função daquele que é o objetivo, como é o caso de regras de associação, árvores de decisão, redes neuronais, classificação Bayesiana, regressão linear, regressão não linear, séries temporais, entre outros (Santos & Ramos, 2017), (Camilo & Silva, 2009), (Han, Pei & Kamber, 2012), (Singh *et al*., 2019).

No Apêndice A poderá ser consultado um estudo efetuado de forma exaustiva a cada uma das restantes componentes anteriormente referidas.

1.4 Ferramentas de BI líderes de mercado

São cada vez mais as empresas a atuarem no mercado do *BI* e a disponibilizarem produtos, serviços e soluções direcionados para as organizações que pretendam obter, através da tecnologia, um auxílio que facilite a sua gestão e orientação, já que o estado atual do mercado tende para uma modificação constante e globalização bastante prolífera (Laruccia, Silva & Chiarelli, 2013).

Assim, surgem as ferramentas de BI, que se definem como *softwares* de recolha e processamento de um grande volume de dados. Oferecem suporte na preparação dos mesmos para posterior análise, dando origem a relatórios, *dashboards* e visualizações intuitivas (Borges *et al*., 2018).

Tendo como base o **Quadrante Mágico de *Gartner*** atualizado ao ano de 2019, que se encontra abaixo na **Figura *3*,** é possível visualizar que as ferramentas de *BI* consideradas líderes de mercado são a ***Tableau*, *QlikView* e *Power BI* (*Microsoft*)**, contribuindo assim para a obtenção de melhores resultados e vantagem competitiva.

Na prática, o Quadrante Mágico de Gartner assenta numa divisão das respetivas ferramentas e dos diversos *players* em quatro grupos, de acordo com as suas avaliações e pontuações obtidas. Assim, o quadrante superior direito é constituído pelos líderes, o quadrante inferior direito engloba os visionários, o quadrante superior esquerdo os desafiadores e o quadrante inferior esquerdo os fornecedores de nicho[4], contudo, só o facto da empresa se inserir em qualquer

[4] Nicho de mercado é um segmento com necessidades particulares que são pouco exploradas

um dos quadrantes já demonstra ser uma referência mundial no que concerne às soluções que disponibiliza (Gartner, s.d.), (Alves, 2019).

Figura 3 – Quadrante Mágico de Gartner, 2020

Fonte: Richardson, Sallam, Schlegel, Kronz, & Sun, 2020

Com vista em se obter uma melhor compreensão das características, potencialidades e principais diferenças de cada uma das três ferramentas líderes, segue-se, de forma sucinta, uma análise comparativa face ao vasto leque de fatores em análise.

Relativamente à ***Tableau***, comprova-se que esta ferramenta lidera o mercado no que diz respeito à visualização de dados, tendo em conta que dispõe de painéis personalizados, extensas conexões de fontes de dados, facilidade na sua utilização por técnicos especializados e não especializados (Castro & Silva, s.d.) aliado a um *software* que se destaca pela sua robustez (Borges *et al*., 2018).

Já o ***QlikView*** destaca-se na sua capacidade analítica, por deter não só os melhores recursos analíticos, mas também por disponibilizar uma análise flexível que permite versatilidade, personalização e atualização aos seus utilizadores (Castro & Silva, s.d.).

Contudo, o ***Power BI*** supera as restantes ferramentas no que respeita à tomada de decisão, já que fornece uma extensa fonte de *know-how* dos seus parceiros inatingível pelos concorrentes (Castro & Silva, s.d.). Em relação à sua utilização, a sua *interface* simples e intuitiva, assume-se como uma vantagem, ganhando especial relevo quando se refere à sua capacidade de integração de grandes volumes de dados provenientes de diversas fontes (Borges *et al*., 2018), (Castro & Silva, s.d.).

É, contudo, importante referir que a ideia da existência de uma ferramenta melhor do que outra é utópica, já que cada uma apresenta benefícios e desvantagens. Claramente que a melhor ferramenta a ser implementada por uma qualquer organização é sem dúvida aquela que melhor se adeque e adapte àquela que é a realidade e contexto da mesma, dando a possibilidade e flexibilidade proporcionais às suas mudanças constantes. Isto é, é de extrema importância que essa ferramenta disponha de recursos e meios necessários à satisfação das necessidades da organização, que seja simples e eficaz na sua implementação, utilização e aprendizagem, em prol de se obterem os melhores *outputs* possíveis.

1.5 Benefícios dos sistemas de BI nas organizações

Qualquer empresa ou organização tem ao seu dispor o usufruto das mais variadas potencialidades, ferramentas e mais valias inerentes aos sistemas de BI, uma vez que são diversas as vantagens e benefícios associados ao uso e adoção de sistemas desta natureza.

Estas ferramentas contribuem para processos de decisão mais sustentados e contextualizados, resultantes de informação fidedigna, oportuna, consistente, segura, precisa e de qualidade proveniente do seu rigor e da sua rápida obtenção de informação perto de tempo real. Tal, permite às organizações identificar as fraquezas e forças de forma expedita, potenciando uma melhor adaptação ao mercado, concedendo-lhes assim, vantagem competitiva (Laruccia *et al*., 2013), (Antonelli, 2009).

Paralelamente, os sistemas de BI contribuem para melhorias de desempenho a nível do negócio e consequentemente a nível organizacional, já que permitem monitorizar desempenhos anteriores e correntes, efetuando análises, projeções e adaptações para o futuro (Santos & Ramos, 2017), (Laruccia *et al.*, 2013), (Antonelli, 2009).

Um aspeto importante a ser salientado é o auxílio prestado ao nível dos resultados de uma organização, ajudando a orientar a estratégia da mesma de acordo com o core e definição rigorosa das metas da própria, potenciando resultados mais próximos dos desejados e permitindo um acompanhamento da situação real (Santos & Ramos, 2017), (Laruccia *et al*., 2013).

Enfatiza-se a criatividade organizacional associada à capacidade de inovação, uma vez que estes sistemas constituem a rampa de lançamento de novas ideias, produtos ou serviços que levam a uma melhor adaptação por parte das organizações ao mercado em constante mudança (Santos & Ramos, 2017), (Laruccia *et al*., 2013).

É de ressalvar que os objetivos principais da adoção deste tipo de sistemas se prendem com a geração e maximização do lucro, reduzindo custos nomeadamente custos de *software*, custos com administração e suporte, custos na avaliação de projetos, entre outros tendo em conta que se pretende que todos os processos inerentes a uma dada organização sejam executados de forma otimizada, evitando desperdícios (Laruccia *et al*., 2013), (Antonelli, 2009).

Claro está que o sucesso deste tipo de sistemas só é possível quando se estabelece um alinhamento perfeito entre a gestão estratégica, tática e operacional bem como a existência de colaboração entre toda a equipa e boa comunicação ao nível da organização (Santos & Ramos, 2017), (Antonelli, 2009).

A sua adesão parte de organizações que procuram alavancar os seus negócios com o desejo de se tornarem mais competitivas no mercado, recorrendo para isso a sistemas diferenciadores que atuem como suporte e apoio às suas decisões. A sua adoção ainda constitui um desafio para muitas organizações, embora nos últimos anos o mercado de BI tenha sofrido algumas alterações, nomeadamente em relação ao aumento dos seus produtos e serviços oferecidos, assim como a sua adesão que também tem vindo a aumentar (Chaudhuri, Dayal, & Narasayya, 2011). Tal ocorre devido sobretudo ao valor inigualável que estes sistemas acrescentam, na medida em que permite facilitar vários processos, abrindo caminho para uma era mais tecnológica, menos intuitiva e mais adequada a cada contexto organizacional.

1.6 Business Analytics

O *Business Analytics* (BA) corresponde à aplicação de ferramentas, técnicas e princípios de análise a problemas de negócio complexos, com o objetivo de alavancar e extrair valor a partir de dados (Acito & Katri, 2014), (Delen, 2015).

O termo *Analytics* refere-se à descoberta de padrões significativos presentes nos dados, fazendo uso de modelos matemáticos sofisticados. Constitui assim, uma diversidade de métodos, tecnologias e ferramentas associadas para gerar conhecimento e *insights*, de forma a resolver problemas complexos e contribuir para tomadas de decisão mais rápidas e de maior qualidade. Facilita, assim, a realização dos objetivos de negócio, otimizando os seus processos com o

propósito de melhorar o seu desempenho (Delen, 2015). Divide-se em três níveis hierárquicos: *Descriptive Analytics*, *Predictive Analytics* e *Prescriptive Analytics*, aumentando o nível de sofisticação e complexidade do primeiro para o último. Contudo, não existe uma clara separação, podendo um negócio estar presente nos três níveis em simultâneo (Delen, 2015). O *Descriptive Analytics* corresponde ao diagnóstico da situação atual (Silva, 2017), dedicando-se sobretudo ao *Business Reporting* e, portanto, à criação de relatórios que respondam a questões como "O que aconteceu?", "O que está a acontecer?" (Delen, 2015). É comummente chamado BI. Como já foi referido no início deste capítulo, o BI é uma das mais populares tendências tecnológicas para sistemas de informação criada para dar suporte aos responsáveis pelas tomadas de decisão. É o ponto de partida para a entrada no mundo do *Analytics*, abrindo caminho para análises mais sofisticadas. Os níveis *Predictive* e *Prescriptive Analytics* são então considerados como *advanced analytics* (Delen, 2015). O *Predictive Analytics* tem como objetivo responder à questão "O que irá acontecer no futuro?", fazendo projeções futuras e utilizando para esse fim, métodos de previsão causais ou séries temporais (Silva, 2017) que será objeto de estudo na subsecção seguinte. O terceiro e último nível diz respeito a um nível estratégico, que pretende apurar de que forma a organização deverá agir futuramente (Delen, 2015) e, para tal, elabora modelos de otimização, simulação e técnicas de modelos de decisão baseadas em heurísticas que apoiem e guiem o processo de decisão (Silva, 2017).

É a combinação entre a necessidade de competências de análise de dados (BA) com a necessidade de tecnologia ao nível de conexão de dados, armazenamento e tratamento dos mesmos (BI) que se dá a criação de valor e de vantagem competitiva (Silva, 2017).

1.7 Análise Preditiva

1.7.1 Conceito

A análise preditiva é uma tarefa de natureza estatística que assenta na descoberta e extração de padrões, relacionamentos e informação a partir de dados históricos e atuais, bem como de todos os *inputs* que possam influenciar este tipo de análise, com o principal objetivo de prever, extrapolando tendências e comportamentos futuros, tentando, desta forma, antecipar cenários. Combina técnicas estatísticas, técnicas de *Data Mining*, *Machine Learning* (ML) e Inteligência Artificial (IA). Fornece essencialmente um bom suporte a tomadas de decisão devidamente sustentadas, desde que as previsões sejam válidas e suficientemente precisas. Pode desta forma, ainda, ser entendida como um pilar para efeitos de planeamento e definição de objetivos estratégicos de curto, médio ou longo prazo, dependendo do horizonte temporal requerido, cujas aplicações se estendem por diversos domínios (Hyndman & Athanasopoulos, 2018), (Gandomi & Haider, 2014), (Kumar & Garg, 2018).

Para esse fim, é necessário que as organizações, em conjunto com os analistas, desenvolvam e implementem um processo de análise preditiva à imagem do contexto onde operam e em função dos seus objetivos. Numa primeira instância, deverão ser identificadas as necessidades adjacentes à previsão. Em seguida, é imperativo que se determine o horizonte temporal (curto, médio ou longo prazo). Depois, há que selecionar os dados que servirão de base à previsão, assim como todos os indicadores considerados importantes e, por último, selecionar o método de previsão que melhor se adeque, sendo que este estará sujeito a uma avaliação e se necessário a uma redefinição do mesmo (Hyndman & Athanasopoulos, 2018), (Adam & Ebert, 1991), (Kumar & Garg, 2018).

1.7.2 Métodos e tipologia de dados

Os métodos de análise preditiva podem ser divididos em dois grupos, **qualitativos** e **quantitativos**, em função dos dados que se encontram disponíveis para análise (Hyndman & Athanapoulos, 2018). Os primeiros aplicam-se na ausência de dados históricos ou, quando na existência de dados disponíveis, estes não apresentam qualquer consistência ou relevância para a previsão, sendo esta bastante subjetiva, baseada em intuições, julgamentos e opiniões. Mais concretamente, os métodos qualitativos recolhem informação e dados de duas fontes principais: interna e externa à organização. A recolha interna, faz uso de opiniões recolhidas de vendedores, gestores e estimativas de clientes. A recolha externa é feita recorrendo a comparações com valores históricos e a uma combinação de todos os possíveis impactos externos. Por fim, ainda se inclui nesta categoria o método de Delphi, posto em prática por grupos estruturados e especializados para a conceção de previsões, sendo que, por essa razão, se assuma que as suas previsões sejam consideradas mais precisas quando comparadas com aquelas elaboradas por iniciativas individuais (Hyndman & Athanapoulos, 2018), (Evans, 2016).

Os **métodos quantitativos** implementam-se quando se satisfazem duas condições em simultâneo, isto é, se por um lado, existirem dados históricos numéricos disponíveis e, por outro, for legítimo assumir que determinadas situações e padrões passados se irão repetir no futuro. Os métodos quantitativos ainda se podem dividir em **métodos causais** e **métodos de séries temporais**. Os **métodos causais** referem-se a modelos de regressão, onde, se assume à priori que exista uma relação de causa-efeito entre as variáveis. Significa isto que, a variável que se pretende prever (também designada por variável dependente ou explicada) é explicada a partir do comportamento e variações de outras variáveis previamente conhecidas (variáveis independentes ou explicativas) (Hyndman & Athanapoulos, 2018), (Evans, 2016). O seu objetivo extravasa as previsões, numa

tentativa de explicar fenómenos de negócio e aumentar a compreensão de relações subjacentes e estabelecidas entre determinadas variáveis (Evans, 2016).

Por sua vez, os **métodos de séries temporais** são aplicados em dados de séries temporais. Dados de séries temporais são um conjunto ordenado e sequencial de observações referentes a uma dada variável em intervalos de tempo regulares (mensalmente, diariamente, de dois em dois meses, etc.). A previsão em séries temporais tem como objetivo prever ocorrências futuras tendo como base um modelo de análise ao histórico da variável a prever, partindo do pressuposto de que o histórico se repetirá no futuro com algumas semelhanças. Para tal, esse modelo tenta captar as componentes e padrões fundamentais presentes nos dados, como tendências, sazonalidade, ciclos ou irregularidades com o propósito de, combinando todos estes *inputs*, tentar projetar o seu comportamento futuro melhorando, desta forma, o processo de tomada de decisões futuras (Singh *et al*., 2019), (Hyndman & Athanapoulos, 2018), (Evans, 2016), (Daitan, 2019). A **tendência** é um indicador do comportamento global dos dados, caracterizando-se por um aumento ou decréscimo constante e gradual nos valores da série temporal ao longo do tempo. A **sazonalidade** define-se por flutuações positivas ou negativas nos dados, que ocorrem num período específico de um dado ciclo, que se repete de forma padronizada e regular para outros ciclos. Exemplo: comportamento particular de vendas que se repete anualmente ou mensalmente ou diariamente de forma semelhante. Por último, as **irregularidades** (comummente designadas por erros) tal como o nome indica, são alterações irregulares e imprevisíveis nos dados ao longo do tempo, sem ditarem qualquer tipo de padrão ou explicação específicos (Singh *et al*., 2019), (Daitan, 2019).

A decomposição e descrição destas componentes pode ser feita através de duas abordagens: aditiva e multiplicativa. Na aditiva, o modelo ilustra-se da seguinte forma: Y = T + S + C + I, onde o Y representa a variável a prever e o T, S, C e I representam a componente de tendência, sazonalidade, ciclos e

irregularidades respetivamente. O modelo aditivo assume que estas componentes são independentes entre si e, por isso, não são influenciadas por nenhuma outra, daí advém a soma. Numa abordagem multiplicativa, o modelo ilustra-se da seguinte forma: Y = T * S * C * I, onde as componentes apresentam o mesmo significado, porém, neste tipo de modelo, pressupõe-se que as componentes não estão isentas de independência total, podendo até influenciar-se mutuamente (Singh *et al*., 2019).

Dentro dos **métodos de séries temporais**, encontram-se diferentes modelos que se podem implementar de acordo com as componentes e características visíveis e existentes nos dados, podendo dividir-se em ***Moving Averages*** e ***Exponential Smoothing*** (Singh *et al*., 2019), (Evans, 2016). Este último será o modelo levado a cabo na demonstração prática, e, por isso, será objeto de explicação detalhada.

Os modelos *Exponential Smoothing* constituem uma técnica de previsão assente em médias ponderadas de observações do passado, onde os dados são suavizados exponencialmente, atribuindo pesos exponencialmente crescentes ou decrescentes em função destes serem mais ou menos atuais, respetivamente. Isto quer dizer que se atribui um maior peso e, por isso, maior importância aos dados, à medida que as observações se tornam mais recentes e atuais. Da mesma forma que se atribui um peso decrescente à medida que os dados se tornam mais antigos (Singh *et al*., 2019), (Hyndman & Athanapoulos, 2018). As previsões por estes modelos produzidas são conseguidas de forma ágil e rápida sendo consideradas bastante viáveis (Hyndman & Athanapoulos, 2018). Este tipo de modelos divide-se em ***Single*** ou ***Simple Exponential Smoothing (SES)***, ***Double*** ou ***Holt's Exponential Smoothing*** e ***Triple*** ou ***Holt-Winters Exponential Smoothing*** cuja escolha do modelo mais adequado dependerá das componentes existentes nos dados e na forma como estas são tratadas pelo modelo, se de forma aditiva ou multiplicativa (Hyndman & Athanapoulos, 2018), (Singh *et al*., 2019).

Assim, o *Simple Exponential smoothing* é um modelo particular de um *Moving Averages* ponderado, onde as observações mais antigas não se perdem, contudo, é atribuído um peso exponencialmente menor à medida que os dados se tornam mais antigos (Evans, 2016), (Singh *et al*., 2019). Utiliza-se nos casos em que os dados não apresentam um padrão de tendência ou sazonalidade claros, existindo apenas uma equação de alisamento, para St, que corresponde aos valores alisados da média das observações no momento t, e, consequentemente apenas uma constante de alisamento, α, que diz respeito ao fator de alisamento das observações, variando entre 0 e 1 (Hyndman & Athanapoulos, 2018), (Singh *et al*., 2019), (Brownlee, 2020), (Evans, 2016). Valores de α perto de 1 significa que o modelo privilegia as observações mais recentes em detrimento das passadas, e por isso, representam um peso maior, enquanto que, por outro lado, quando se assumem valores de α perto de zero, acontece o inverso e por isso os acontecimentos mais passados influenciam o modelo de previsão (Evans, 2016). Contempla ainda uma equação para o cálculo de valores previstos para os períodos desconhecidos que serão todos iguais ao valor suavizado para o último período observado (Hyndman & Athanapoulos, 2018) (Evans, 2016).

O *Double* ou *Holt Exponential Smoothing* é uma extensão do *Simple Exponential Smoothing*, utilizado em casos em que os dados apresentam padrões de tendência clara, de aumento ou diminuição, podendo esta, refletir-se de duas formas distintas, linear e por isso, o modelo a aplicar deverá ser o aditivo, ou exponencial cujo melhor modelo a implementar será o multiplicativo. No caso do *Holt Model*, assume-se que a tendência seja linear (Brownlee, 2020). Este, inclui duas equações de alisamento, uma para os valores de St, já previamente explicados e outra para os valores de Tt, que representa uma estimativa da tendência no momento t (Singh *et al*., 2019), (Hyndman & Athanapoulos, 2018), (Evans, 2016), (Brownlee, 2020). Associadas às equações, estão duas constantes de alisamento: α e β, que correspondem, respetivamente, ao parâmetro de alisamento dos dados e ao

parâmetro de alisamento da tendência, variando ambas entre 0 e 1 (Singh *et al*., 2019), (Hyndman & Athanapoulos, 2018), (Evans, 2016), (Brownlee, 2020). Assim, a determinação do valor de β vai ter em conta a seleção prévia do α, onde, se fixará o valor deste e se alterará o valor do β. Quando este assume um valor igual a zero, significa que a tendência ao longo do tempo é constante. Inclui, ainda, uma equação de cálculo de valores previstos (Evans, 2016).

Por último, o *Triple* ou *Holt-Winters Exponential Smoothing* é uma extensão do modelo *Holt Exponential Smoothing* aplicado em situações cujos dados apresentam padrões claros de tendência e sazonalidade em simultâneo, tendo por isso, três equações de alisamento: uma para os valores de St, outra para os valores de Tt e ainda outra para os valores de It, onde as duas primeiras apresentam o mesmo significado dos modelos anteriores, e a It representa a sequência de fatores de correção de sazonalidade. Dispõe ainda de três constantes de alisamento: α, β e γ, sendo que as duas primeiras constantes de alisamento assumem o mesmo significado dos modelos anteriormente descritos, e γ, a constante de alisamento correspondente à sazonalidade, variando entre 0 e $1-\alpha$. O valor destas três constantes deve ser estimado de tal forma que minimize as métricas de erro escolhidas (Singh *et al*., 2019), (Hyndman & Athanapoulos, 2018) (Evans, 2016), (Brownlee, 2020). Este modelo pode ainda dividir-se em multiplicativo e aditivo. O primeiro aplica-se a modelos cuja sazonalidade é exponencial, isto é, quando as variações sazonais se alteram de forma proporcional ao longo da série temporal. Por outro lado, quando a sazonalidade se apresenta de forma linear, ou seja, quando as variações sazonais são aproximadamente constantes ao longo das observações, o modelo apropriado é o aditivo (Hyndman & Athanapoulos, 2018), (Brownlee, 2020).

Todas as fórmulas e significado inerentes a cada um dos modelos, podem ser consultados respetivamente no apêndice A.

1.7.3 Métricas de erro

As métricas ou medidas de erro têm como principal propósito avaliar a precisão, qualidade e a fiabilidade das previsões e do desempenho dos modelos de previsão, sustentando a escolha do modelo mais sofisticado, que se caracteriza por produzir previsões mais robustas e com menores níveis de erro (Singh *et al*., 2019), (Hyndman & Athanapoulos, 2018), (Bhattacharjee, 2019), (Adam & Ebert, 1991).

Estas medidas são ainda um excelente indicador a ter como base, na seleção de valores arbitrários, a assumir para constantes de alisamento, uma vez que se se devem assumir constantes de alisamento que minimizem as métricas de erro (Singh *et al*., 2019), (Brownlee, 2020).

No **Quadro 1** encontram-se descritas algumas das métricas de erro mais utilizadas.

Métrica de Erro	**R^2**	**Error**	***Mean Absolute Deviation (MAD)***	***Mean Percentage Error* (MPE)**	***Mean Squared Error* (MSE)**	***Mean Absolute Percentage Error* (MAPE)**
Fórmula	$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$	$Error = y - \hat{y}$	$MAD = \frac{\sum_{i=1}^{n} \lvert y - \hat{y} \rvert}{n}$	$MPE = \frac{100\%}{n} \sum \left(\frac{y - \hat{y}}{y} \right)$	$MSE = \frac{\sum_{i=0}^{n} (y_i - \hat{y}_i)^2}{n}$	$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left\lvert \frac{y_i - \hat{y}_i}{y_i} \right\rvert$
Descrição	Coeficiente que mede a qualidade da reta de regressão. Quanto maior for o coeficiente, melhor está a reta ajustada.	Mede a diferença entre o valor observado e o valor previsto.	Mede o erro em unidades e calcula o erro absoluto médio.	Determina a média calculada dos erros em %.	Calcula a média do quadrado dos erros.	Indica a média em % dos erros.

Quadro 1- Métricas de Erro

Capítulo 2
Microsoft Power BI

2. O Recurso Tecnológico Power BI

2.1 Caracterização global

Alinhado com o capítulo anterior, a extração de conhecimento útil que facilita o processo de tomada de decisões diárias e devidamente fundamentadas vem sucedida de uma análise profunda a todo o conjunto de dados e *inputs* que são fornecidos e disponibilizados aos devidos responsáveis. É então neste sentido que surge o *Microsoft Power BI*, um ecossistema que integra o BI corporativo já existente com o conceito de *self-service BI* (Ferrari & Russo, 2016). Esta tendência visa dar autonomia aos utilizadores que não dispõem da formação especializada que os sistemas analíticos exigiam no passado, para executarem determinadas tarefas que até então estavam assim adstritas ao BI corporativo (Sikes, 2017), (Powell, 2018).

O *Power BI* é um conjunto de serviços de *software*, aplicações e conectores que cooperam de forma integrada para transformar as origens de dados não relacionadas em informações coerentes, visualmente envolventes e intuitivas que sustentam o processo de tomada de decisões (Microsoft, 2019).

Lançada em 2015 pela *Microsoft*, esta ferramenta caracteriza-se ainda pela sua simplicidade de utilização, no que respeita a análises avançadas, conexão e integração de dados provenientes de diferentes fontes que podem ser locais ou em *cloud*, novas visualizações de dados, criação e partilha de relatórios, *dashboards* e *insights* de negócio com terceiros, de forma *online* (Pereira, 2020),

(Sonna, 2018), (Microsoft, 2019), (Ferrari & Russo, 2016). Desta forma, os utilizadores conseguem obter uma visão de 360º relativamente aos indicadores e métricas mais relevantes (KPI) do seu negócio, disponibilizadas num só local, atualizadas em tempo real de forma simples e eficiente, estando disponíveis em todos os dispositivos, desde telemóveis, *tablets*, computadores. Isto permite, às organizações, obter um maior índice de produtividade e eficiência no que concerne ao processo analítico, ao efetuar-se um levantamento interativo e profundo acerca das mesmas e simultaneamente delinear estratégias para os seus objetivos futuros (Portal de Gestão, 2017), (Ferrari & Russo, 2016).

O PBI está disponível em três plataformas principais, nomeadamente o ***Power BI Desktop***, uma aplicação para computadores com *Windows*, o ***Power BI service*** que corresponde a um serviço *online* baseado em SaaS (*Software as a Service*) que permite que todo o processo seja gerido (implementação e acesso) na *cloud* (Microsoft Azure) e o ***Power BI Mobile***, que permite o acesso em aplicações móveis *Android*, iOS e *Windows* (Microsoft, 2019). Estas três ferramentas são, regra geral, utilizadas em simultâneo, tendo início na conexão a fontes de dados, dando origem a relatórios e *dashboards* criados no *Power BI Desktop*. Seguidamente são publicados no *Power BI service* e, por último, partilhados através do *Power BI Mobile* possibilitando aos seus utilizadores ver e interagir com os mesmos. Daqui conclui-se que estes três processos se complementam entre si (Microsoft, 2019).
Adicionalmente, existem outros componentes que auxiliam na criação e partilha de relatórios como o ***Power Query***, uma ferramenta de extração e conexão de dados, provenientes de diversas fontes, responsável pela sua transformação (Rad, 2019), o ***Power Pivot*** que serve essencialmente para modelar os dados, alterando-os de diversas formas, como é o caso de criação de relações entre tabelas, criação de KPI's, alteração da estrutura de dados recorrendo ao DAX (*Data Analysis eXpression*), entre outras funções (Rad, 2019). O DAX é um conjunto de recursos de funções, operadores e constantes que serve para transformar os dados, por meio de fórmulas ou expressões, ajudando a obter novas informações

provenientes dos mesmos (Microsoft, 2019). O ***Power View*****,** é uma componente de exploração, visualização e apresentação de dados interativa bem como o ***Power Map*** que, à sua semelhança constitui uma ferramenta de visualização de dados, mas de natureza geoespacial em 3D (Rad, 2019). O ***Power Q & A*** é um mecanismo de linguagem natural que permite que se façam perguntas e obtenham respostas a partir do conjunto de dados em análise. Atua em conjunto com o *Power View*, originando elementos gráficos (Rad, 2019). O ***Power BI Gateway*** maioritariamente utilizado para assegurar de forma rápida e segura a transferência entre dados no local (dados que não se encontram na *cloud*, comummente designados por *on-premises data*) e vários serviços *cloud* da *Microsoft* e ainda a sincronização de dados dentro e fora do *Power BI (*Microsoft, 2019). ***O Power BI Embedded*** agiliza as funcionalidades do *Power BI* ao adicionar de forma simples elementos visuais, relatórios e *dashboards* às suas aplicações (Microsoft, 2018). O ***Power BI Report Server*** que gera relatórios de BI para empresas que precisam de manter os seus dados e relatórios em servidores locais (Microsoft, 2020) e o ***Power BI Visual Marketplace*** que consiste numa *appSource* bastante útil que contém diversos pacotes visuais e analíticos personalizados, com base em bibliotecas *JavaScript* e *scripts* de linguagem R (Microsoft, 2019).

Atualmente existem três versões do *Power BI* no mercado: o ***Power BI desktop*** (ou *Power BI*) que diz respeito a uma versão gratuita e elementar que se encontra ao dispor de qualquer utilizador (Louzada, 2019). O ***Power BI Pro***, uma versão que, embora não seja gratuita, é passível de ser experimentada isenta de qualquer custo. A subscrição da licença inerente a esta versão é mensal e por utilizador individual. Assemelha-se à versão *Desktop*, exceto no que respeita à partilha de dados, relatórios e *dashboards* que pode ser feita com terceiros de forma privada, desde que estes também disponham de uma licença Pro (Louzada, 2019). O ***Power BI Premium*** destina-se a um ambiente *corporate* onde se analisam quantidades significativas de dados (Microsoft, s.d.), (Wade, 2019), (Tkachuk**,** 2019). Como tal, a subscrição da sua licença traduz-se nas necessidades, por parte

das organizações, em termos de acesso a recursos de computação e armazenamento na *cloud*, isto é, refere-se ao espaço e capacidade de processamento que o seu negócio precisa, optando por implementar o modelo que melhor se enquadre. Ao contrário das versões anteriores, não existe a possibilidade de partilha de relatórios a não ser que se subscreva em paralelo uma licença adicional de PBI (Microsoft, s.d.).

2.2 Componente Desktop

O *Microsoft Power BI* funciona através da conexão a fontes de dados, sendo estas provenientes de várias naturezas e, com o objetivo de disponibilizar uma solução de *BI* para os seus utilizadores. Os dados obtidos provenientes de fontes em nuvem são atualizados automaticamente. Por outro lado, se as pastas de trabalho do Excel ou arquivos do PBI estiverem conectadas a fontes de dados *on-premises*, os utilizadores do *Power BI* devem atualizar ou configurar manualmente uma atualização, de forma a assegurar que os dados utilizados na conceção de *dashboards* e relatórios sejam os mais atuais (Iseminger, 2020).

Seguidamente, os analistas e outros utilizadores podem combinar os dados provenientes dessas fontes (fenómeno designado por modelação), formatando-os através de consultas, podendo assim criar modelos de dados que melhor se adequem aos propósitos requeridos (Iseminger, 2020). Uma vez criados, estes modelos permitem que se gerem visualizações e relatórios que podem ser, por um lado, partilhados como ficheiro *PBI Desktop* (.pbix) ou carregados para o *Power BI Service* sendo este último o método mais conveniente (Iseminger, 2020). Em suma, o *PBI Desktop* é onde ocorre todo o processo anteriormente descrito, de forma mais simplificada e concentrada, em comparação com o que acontecia até então. A utilização combinada deste com o *PBI service,* vem potenciar este

processo, que culmina com a sua partilha através do serviço *online* (Iseminger, 2020).

No que à sua utilização diz respeito, o *PBI Desktop* disponibiliza três vistas distintas, que podem ser selecionadas através dos respetivos ícones, no lado esquerdo da tela. As mesmas, surgem pela ordem exposta abaixo:

- **Relatório**: Nesta vista, é possível a criação de relatórios e elementos visuais (Iseminger, 2019);
- **Dados**: Nesta vista, é possível visualizar todas as tabelas, indicadores e outros dados usados no modelo de dados associado ao relatório, assim como transformar os dados para melhor utilização no modelo do relatório (Iseminger, 2019);
- **Modelo/Relações**: Nesta vista, é possível visualizar todas as tabelas, colunas e relações no modelo de dados, bem como gerir as relações estabelecidas entre as tabelas no mesmo (Iseminger, 2019).

2.3 Utilizadores Power BI

O valor e domínio que o *Power BI* apresenta reside na versatilidade que oferece para servir e melhorar o desempenho de diversas funções dentro de uma dada organização, maximizando a sua eficiência (Blast, 2019). Como tal, e num âmbito empresarial, é vasto o tipo de utilizadores que podem usufruir e tirar partido das funcionalidades desta ferramenta, podendo segmentar-se em três grupos principais, tendo por base o nível de competências e autonomia que apresentam na utilização da mesma, nomeadamente: *Developers – 1st wave, Power BI analysts – 2nd wave* e *End Users – 3rd wave* (Velosio, 2019). **A Figura 4** ilustra esta divisão.

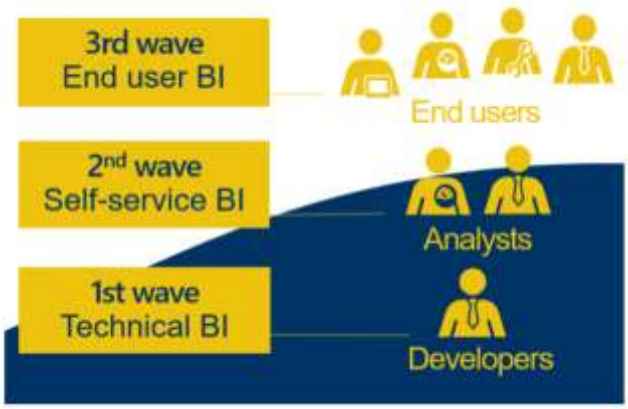

Figura 4 - Tipos de utilizadores de PBI

Fonte: Velosio, 2019[5]

Como a figura evidencia, na base da escala, e mais concretamente no que diz respeito aos *Developers*, estes deverão possuir um nível elevado de *skills* e autonomia no que respeita ao tratamento e gestão de dados, ter a perspicácia de identificar e entender as necessidades e por consequência, as configurações a aplicar aos dados, facilitando assim a tarefa aos *End Users* (Velosio, 2019), (Bouwens, 2018). Adicionalmente, são responsáveis por todos os cálculos e medidas criadas com recurso à linguagem DAX, sendo imperativo deter conhecimentos desta linguagem e igualmente em SQL (Velosio, 2019), (Bouwens, 2018). Por último, este tipo de utilizador deverá ainda revelar grande capacidade criativa, concebendo relatórios apelativos (Velosio, 2019).

No segundo nível encontram-se os *Analysts*, que, por sua vez, deverão ter um conhecimento básico de dados, não necessitando de os trabalhar ou transformar, uma vez que as BD são disponibilizadas já devidamente trabalhadas e atualizadas pelos *Developers*, o que também os isenta, por sua vez, da necessidade de conhecimentos de DAX ou SQL (Velosio, 2019). No entanto, deverão ser conhecedores das relações entre as tabelas aí existentes, conferindo-lhes a

[5] Consultado em https://www.erpsoftwareblog.com/2019/09/powerbi-license-types-developers-analysts-end-users/ a 24/03/2020

possibilidade de criar e editar relatórios (podendo ser relatórios *ad-hoc*), que evidenciem com clareza aquilo que se pretende demonstrar, recorrendo para isso à sua criatividade (Velosio, 2019), (Bouwens, 2018).

Por fim, no topo da figura, estão os *End Users*, utilizadores que, têm um menor nível de autonomia e conhecimento técnico, pelo que a sua utilização do *Power BI* se cinge à navegação na sua *interface*, fazendo uso das várias funcionalidades de que a ferramenta disponibiliza, como filtrar e dividir os dados (Velosio, 2019), (Bouwens, 2018).

2.4 Potencialidades

Como tem vindo a ser referido, o objetivo principal desta ferramenta prende-se sobretudo com a geração de relatórios e *dashboards* como auxílio na tomada de decisões (Pereira, 2020). Neste âmbito, a mesma dispõe de um leque de recursos e funcionalidades que facilitam a vida aos seus utilizadores.

De forma enumerativa, é possível listar as principais potencialidades da ferramenta, nomeadamente (Scardina & Horwitz, s.d.), (Sonna, 2018):

1. Criação de relatórios e *dashboards* interativos;
2. Partilha de várias páginas num só relatório;
3. Possibilidade de customização através do *MarketPlace;*
4. Atualização de dados e análises em tempo real ou à frequência desejada;
5. Análises simples e intuitivas com recurso a visualização de dados;
6. Análises ad hoc;
7. Análises avançadas com recurso ao R;
8. Funções de Inteligência Artificial (IA);
9. Integração com o Microsoft Excel;
10. Criação de cenários futuros;

11. Conexão com milhares de fontes de dados;

12. Partilha de *dashboards* na *web* ou em dispositivos móveis;

13. Integração de contributos parceiros;

14. Integração de outros produtos da Microsoft;

Materializando alguns dos pontos acima escritos, no que diz respeito à geração de relatórios e *dashboards*, estes são habitualmente criados como forma de visualizar e analisar numa base quer seja mensal, anual, diária ou até mesmo em tempo real, a situação global da organização, seja internamente ou mesmo relativamente aos seus concorrentes. Este é um processo de otimização de tempo, na medida em que, após a correta definição dos KPI's, a sua conceção é feita de forma rápida e interativa, que possibilita obter *insights* e uma visão mais clara do negócio, acrescentando-lhe desta forma valor, estando a sua atualização à distância de um "clique" (Malheiro, 2020), (Ludovice, 2017).

O PBI conjuga funcionalidades de IA com o BI, oferecendo diversas funções e recursos nos seus relatórios, melhoradas ao longo do tempo, como é o caso dos *Quick Insights* e *Natural Language Queries* A primeira característica assenta na descoberta automática de padrões, tendências e potenciais relações nos dados, por parte da ferramenta, sob a forma de gráficos de simples interpretação (Powell, 2018), (Ferrari & Russo, 2016). Através da funcionalidade *Natural Language Queries*, o Power BI possibilita questionar ou solicitar diretamente ao sistema aquilo que se pretende visualizar e analisar, sendo que para o efeito a instrução deverá ser feita em inglês. *O output* gerado são elementos visuais que ilustram as análises requeridas (Ferrari & Russo, 2016), (Rad, 2019).

Relativamente à visualização de dados, o PBI contribui para uma gestão de análise personalizada, ao disponibilizar um conjunto diferenciador de tipos de visualização, potenciando e facilitando o processo de tomada de decisões, na medida em que as disposições gráficas de um dado conjunto de dados tornam as

decisões e conclusões mais intuitivas e rápidas (Malheiro, 2020), (Ludovice, 2017).

No que diz respeito a análises avançadas, o PBI trabalha com recurso à linguagem R. Esta é uma linguagem de programação bastante complexa utilizada por estatísticos, *data scientists* e *data analysts,* sendo considerada a linguagem estatística do mundo (Microsoft, s.d.). Quando aplicada em conjunto com o PBI, enriquece e eleva as suas análises e geração de *insights*. A sua integração no PBI é feita de uma forma bastante fácil, uma vez que é suficiente ter a versão correta do R instalada no computador já que, todos os pacotes necessários à execução de instruções são automaticamente instalados, bem como os *scripts* necessários são automaticamente executados, sem haver a necessidade de programação.

No mesmo seguimento, o PBI inclui e integra contributos de parceiros ou recorre a outros produtos criados pela *Microsoft* como é o caso do *Excel* ou do *Azure*, sendo este último explicitado na secção 2.5.2 que se segue. O principal objetivo é o de tornar a ferramenta mais potente, fiável e sem grande margem de erro.

A partir daqui, afere-se que com esta ferramenta, os processos de análise de dados se tornam mais eficientes, refletindo-se num maior nível de produtividade, contribuindo para melhorar a *performance* geral da organização (Malheiro, 2020), (Ludovice, 2017).

Um aspeto a ter em consideração na sua adesão é o seu fácil acesso, que se justifica por um lado, derivado à sua condição gratuita, pelo menos da versão mais elementar e por outro com a sua acessibilidade, possível a partir de qualquer local.

Adicionalmente, a sua partilha é de fácil expansão, através de diversas formas, recorrendo a computadores pessoais ou dispositivos móveis e a sua facilidade de importação e exportação de dados é igualmente motivo de destaque (Ludovice, 2017), (Pereira, 2020), (Nabler, s.d.).

Salienta-se, ainda, a diversidade de áreas que podem beneficiar com ferramenta, desde a área Financeira, Engenharia, TI, *Marketing*, a Área Comercial, entre outras.

Como forma conclusiva, é importante ressalvar que, sendo o PBI uma ferramenta pertencente à *Microsoft*, existe a possibilidade de ajuda, esclarecimentos e de aprendizagem, a partir da sua rede com outros clientes no que respeita a dicas e exercícios (Nabler, s.d.). Paralelamente, o seu poder encontra-se, além do que foi referido, no grande investimento de inovação que a *Microsoft* tem vindo a fazer em relação a Serviços Cognitivos, *Machine Learning*, IA, entre outros, que objetiva alcançar utilizadores com conhecimentos e competências mínimas em *Data Science*, desenvolvendo esta ferramenta de utilização intuitiva (Microsoft, 2019).

2.5 Recursos complementares ao Power BI

2.5.1 Integração de contributos de parceiros da Microsoft

A Microsoft, com o principal propósito de melhorar a ferramenta PBI, tornando-a desta forma mais completa, tem adotado a estratégia de conjugação e integração de recursos e soluções provenientes de vários parceiros, em contextos variados.

No que diz respeito ao acesso de dados, o **Azure Event Hub**, **Pub Hub** e o **Temboo** constituem alguns dos fornecedores de dados de *streaming*, que permite que o PBI aceda aos dados em tempo real independentemente da sua fonte ou origem, sendo uma mais valia para os seus utilizadores.

A solução **Alteryx Starter Kit for Microsoft** ilustra o cenário num contexto de visualização de dados, uma vez que a fase crucial que antecede a visualização de dados é a sua preparação que, tem a si associada um peso relevante em termos

de tempo, revelando-se bastante ineficiente em alguns casos. É através da parceria com esta solução integrada que os analistas otimizam o processo de preparação de dados, o que lhes permite passarem rapidamente para a fase de criação de visualizações bastante avançadas de dados com o *Power BI* (Martinez, 2016).

No mesmo seguimento, já num contexto de previsão, o *Power BI* recorre ao pacote **Prevedere** com vista em facilitar a previsão face à *performance* futura das organizações. O mesmo dispõe de tecnologia mais recente na área da computação em nuvem e potencia informações relacionadas com previsões financeiras, como é o caso de informações face aos principais mercados em crescimento, bem como de dados económicos específicos do setor, dados meteorológicos que influenciam a compra de determinados produtos e serviços e das abordagens relevantes para cada negócio (Palmer, 2016). Adicionalmente, os relatórios interativos permitem analisar eventuais externalidades que possam influenciar o negócio, mais concretamente no que ao comportamento do consumidor diz respeito (Palmer, 2016).

2.5.2 Integração de contributos da Microsoft

Expandindo os seus horizontes, o PBI segue uma política de incorporação de contributos de outros produtos provenientes da *Microsoft*, tornando-se uma ferramenta *self-service* de BI bastante versátil, completa, poderosa e única no mercado em diversos contextos.

Enquadrado na análise de dados, o PBI permite aprofundar e aceder a dados em tempo real através da integração do **Microsoft Azure Analytics** e o **Azure Machine Learning.**

No que respeita à análise preditiva, num cenário que conjuga as funcionalidades do PBI numa versão *Premium*, que podem ser consultadas no

quadro do Apêndice B, com as potencialidades do *Azure Machine Learning*, há conceitos essenciais de IA que são úteis à sua compreensão. Como já foi referido anteriormente, o PBI tem investido bastante em características e funções ligadas à IA, ao disponibilizar duas ferramentas poderosas ao nível da mesma (Ulag, 2019):

I. **Azure Cognitive Services:** constituem modelos de *machine learning* pré-testados utilizados em aplicações inteligentes, onde os analistas conseguem extrair *insights* a partir de imagens ao detetarem objetos relevantes. Os campos textuais, e, portanto, os dados não estruturados, como *feedbacks* de clientes, e-mails, entre outros, também são alvo de análise extraindo-se frases-chave ou até mesmo aferir se dado comentário é positivo ou negativo. Tal é possível de se realizar através de relatórios interativos em PBI (Ulag, 2019).

II. **Azure Machine Learning (AML):** por definição, é uma plataforma poderosa onde os *data scientists* conseguem desenvolver modelos de *machine learning* que são facilmente partilhados e utilizados por analistas. Ou seja, o PBI descobre automaticamente quais os modelos pré-concebidos que estão disponíveis para um dado analista, disponibilizando um ponto intuitivo de forma a invocá-lo. Tal, permite que os analistas colaborem com os *data scientists*, além de conseguirem visualizar e utilizar informações desse modelo nos seus relatórios (Ulag, 2019).

De uma forma simplista, a análise preditiva explora factos, dados atuais e históricos que servem de base ao processo de previsão de eventos futuros ou desconhecidos, utilizando para esse fim, padrões captados nesses dados históricos e transacionais (Parashar, 2019). Este tipo de análise surge na emergência de ajudar os utilizadores a desenvolver modelos preditivos de forma

rápida e fiável, a partir de diferentes dados, permitindo que as empresas consigam tomar decisões baseadas em todos os aspetos do seu negócio (Microsoft, s.d.). Para tal, é necessário combinar três serviços principais: o AML, o PBI e um aliado fundamental integrado no PBI, o R, que vai atuar como uma ponte de ligação entre os dois serviços (Lucznik, 2016). Por um lado, com o AML *Studio*, os utilizadores conseguem desenvolver modelos preditivos de forma rápida, ao arrastar, soltar e conectar os módulos de dados e por outro, através do PBI, conseguem visualizar os resultados provenientes dos algoritmos de ML (Microsoft, s.d.).

Adicionalmente, é possível utilizar o *Machine Learning* (ML) de forma automática (AutoML) integrado no Power BI, que, facilita todo o processo de desenvolvimento de um modelo, ao treiná-lo e aplicando as etapas iterativas necessárias de forma automática (AutoML, 2020). Este cenário aplica-se a analistas de negócio, *data professionals* e *developers* que não disponham de *background* ou de conhecimentos profundos em áreas relacionadas, consigam construir modelos eficientes, produtivos e com elevada qualidade (AutoML, 2020), (Parashar, 2019).

Capítulo 3
Design Metodológico

3. Métodos de análise de dados e conceção de LO

3.1 Métodos de análise de dados

A questão de investigação que um investigador formula procura exprimir o que pretende conhecer, descobrir, compreender ou aprofundar. Subjacente à questão de investigação está a explicitação de objetivos que realisticamente clarificam o rumo que orienta a procura de resposta. A questão formulada neste projeto "Qual o potencial do *Power BI* para a análise preditiva"? convoca uma ação concreta de estudo de uma aplicação de *software* num contexto que replique os objetivos formulados.

Neste enquadramento, o paradigma de investigação muitas vezes adotado na área dos sistemas de informação, cujo âmbito de estudo, normalmente se cruza com o comportamento humano e o organizacional é o *Design Science Research* (Hevner & March & Park & Ram, 2004). Trata-se de um quadro concetual que enquadra a procura de um guia claro para definir ideias, práticas, técnicas, produtos, análises, implementação ou a gestão de sistemas de informação de forma inovadora ou mais eficiente.

Nesta procura de resposta através de um guia explícito para realizar uma análise específica de *software* de manipulação analítica de dados, identificam-se as melhores técnicas para o efeito: CRISP-DM, SEMMA e KDD. São três processos adotados em projetos de *Data Mining* que identificam etapas precisas para o bom

êxito deste complexo exercício, assegurando suporte e apoio a projetos desta natureza. Após a sua descrição e análise, a sua aplicação será explicada pela realização de *Learning Objects* elementares para facilitar a partilha do conhecimento.

3.1.1 CRISP-DM

O **CRISP-DM** (*Cross Industry Standard Process for Data Mining*) proporciona uma visão geral do ciclo de vida de um projeto de *Data Mining*, sendo este constituído por seis etapas, que se encontram explicadas e posteriormente ilustradas na **Figura 5**. A sequência destas não é linear, o que significa que a sua ordem é arbitrária, quando o processo assim o exigir. As setas presentes na figura simbolizam as principais relações de dependência entre as mesmas. Este modelo envolve, assim, a necessidade de compreensão dos objetivos do negócio, para que as análises desenvolvidas sirvam de facto o desenvolvimento do mesmo (Azevedo & Santos, 2008), (Chapman *et al*., 2000). Este processo é considerado um dos mais utilizados, não só por demonstrar a versatilidade de poder ser aplicado a qualquer tipo de negócio, mas também por não estar dependente de nenhuma ferramenta que dite a sua execução, sendo bastante completo e bem documentado (Vasconcellos, 2017), (Azevedo & Santos, 2008).

(1) Compreensão do negócio: é nesta fase inicial que se pretende conhecer melhor a perspetiva de negócio, os problemas a solucionar e os seus objetivos principais, para se efetuar a sua conversão em objetivos de *Data Mining* e consequente planeamento para se atingir o pretendido (Santos & Ramos, 2017), (Azevedo & Santos, 2008), (Fávero, 2019), (Chapman *et al*., 2000).

(2) Compreensão dos dados: esta etapa foca-se, numa abordagem inicial, numa recolha e compreensão de dados. Seguidamente, centra-se no levantamento de possíveis problemas de qualidade nestes, como deteção

de anomalias, dados em falta ou redundâncias. É neste ponto que se identificam os primeiros padrões, como possíveis relações de interdependência entre variáveis e a identificação dos conjuntos de dados relevantes e interessantes que serão seguidamente analisados, em vista à identificação de conhecimento camuflado nos mesmos (Santos & Ramos, 2017), (Azevedo & Santos, 2008), (Fávero, 2019), (Chapman *et al*., 2000).

(3) Preparação dos dados: esta fase cobre todo o conjunto de atividades necessárias e levadas a cabo tendo em vista a obtenção de um conjunto de dados final e devidamente consistente, pronto para ser utilizado para análise. Engloba tarefas de limpeza, transformação, redução ou integração de dados, podendo ser implementadas as vezes que forem requeridas (Santos & Ramos, 2017), (Azevedo & Santos, 2008), (Fávero, 2019), (Chapman *et al*., 2000).

(4) Modelação: esta é a etapa de construção e estimação do modelo ou de vários modelos, onde para isso, podem ser aplicadas diferentes técnicas de modelação de dados para o mesmo tipo de problema de *Data Mining*, como é o caso dos algoritmos de *Data Mining* implementados de acordo com os objetivos de análise. Estes, podem ser ajustados ao nível dos seus parâmetros para se obter valores ótimos. Como cada técnica apresenta diferentes requisitos de formatos de dados, pode ser necessário voltar à fase de preparação dos mesmos (Santos & Ramos, 2017), (Azevedo & Santos, 2008), (Vasconcellos, 2017), (Fávero, 2019), (Chapman *et al*., 2000).

(5) Avaliação: esta é uma fase que se dedica à avaliação da qualidade que o(s) modelo(s) obtidos na fase anterior apresentam. É crucial a participação de gestores do negócio, bem como de especialistas estatísticos e analistas, para que se cruzem diversas avaliações aos modelos. Estas, poderão basear-se em métricas de erro, confiança estatística, entre outros indicadores que permitam aferir se os modelos satisfazem todas as condições para a obtenção dos objetivos de negócio. No final, e atendendo a todos os

critérios a serem considerados, é necessário tomar a decisão destes serem ou não utilizados na organização (Santos & Ramos, 2017), (Azevedo & Santos, 2008), (Fávero, 2019), (Chapman *et al*., 2000).

(6) Desenvolvimento: Por último, é nesta fase, que após a modelação e análise dos seus *outputs* se dá início à produção e implementação do modelo, sendo necessário que a informação explícita nos modelos seja difundida à organização de forma transversal, percetível e num formato que permita à mesma a sua correta utilização (Santos & Ramos, 2017), (Azevedo & Santos, 2008), (Fávero, 2019), (Chapman *et al*., 2000). É importante ressalvar que é necessário efetuar uma monitorização periódica aos resultados, bem como efetuar qualquer alteração ou adaptação para melhoria do modelo (Vasconcellos, 2017).

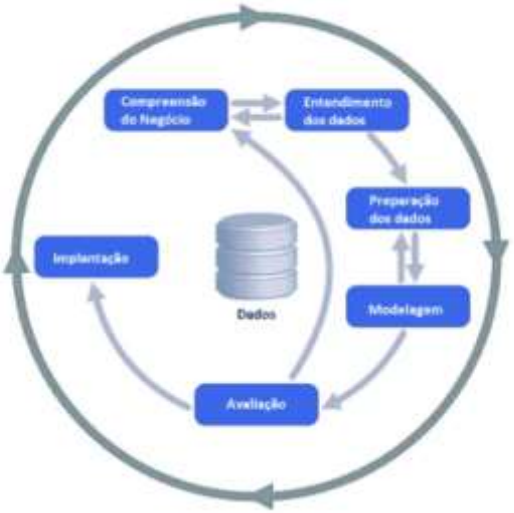

Figura 5 - Etapas do ciclo CRISP-DM

Fonte: Jorge Ramos, 2016[6]

3.1.2 SEMMA

O **SEMMA (**Sample, Explore, Modify, Model e Assess), sendo o seu acrónimo um reflexo das cinco fases inerentes à condução de um projeto de *Data Mining*,

[6] Consultado em https://www.researchgate.net/figure/Figura-16-Fases-do-CRISP-DM_fig11_312605379 a 02/04/2020

foi criado pelo *SAS Institute* e consequentemente está dependente de uma ferramenta específica: *SAS Enterprise Miner* (Santos & Ramos, 2017).

É uma organização lógica de um conjunto de ferramentas funcionais do *SAS Enterprise Miner*, de maneira a executar as principais tarefas de *Data Mining*. O *Enterprise Miner* pode ser usado como parte iterativa de uma qualquer metodologia de *Data Mining* adotada pelo cliente (SAS Institute, 2006).

É bastante fácil de entender e aplicar, conseguindo desenvolver soluções para os problemas propostos e delinear objetivos de negócio de forma expedita e simples (Azevedo & Santos, 2008). Apresenta semelhanças com o CRISP-DM ao focar-se essencialmente na criação do modelo de dados, mas desconsidera as questões e objetivos de negócio (Vasconcellos, 2017), (SAS Institute, 2006).

As respetivas fases são brevemente explicadas abaixo e devidamente ilustradas pela **Figura 6**.

(1) Amostragem: é nesta fase que se dá a redução da BD com a extração de uma amostra pequena de fácil manipulação, mas que seja suficientemente significativa e representativa da BD geral, contendo toda a informação e padrões considerados relevantes e cruciais dos dados (Azevedo & Santos, 2008), (SAS Institute, 2006).

(2) Exploração: esta fase dedica-se à exploração dos dados na tentativa de se anteciparem tendências e anomalias com o objetivo de se extrair conhecimento e gerar ideias (Azevedo & Santos, 2008), (SAS Institute, 2006).

(3) Modificação: esta etapa consiste na modificação dos dados ao criar, selecionar e transformar variáveis, baseando-se nas descobertas na fase anterior. É necessário identificar *outliers* e reduzir a amostra, desconsiderando variáveis irrelevantes, focando-se nos atributos e dados realmente importantes (Azevedo & Santos, 2008), (SAS Institute, 2006).

(4) Modelação: esta etapa ocorre quando se dá permissão ao *software* para que, de forma automática, consiga encontrar um modelo fiável que melhor se

adeque e reflita os padrões/relacionamentos encontrados nos dados. Para tal, são utilizadas algumas técnicas de modelação como redes neuronais, modelos baseados em árvores de decisão, modelos logísticos, análise de séries temporais, entre outros (Santos & Ramos, 2017), (SAS Institute, 2006).

(5) Avaliação: esta última fase foca-se em avaliar a utilidade e fiabilidade dos resultados provenientes do modelo utilizado, de forma a aferir a qualidade do desempenho do mesmo para o propósito (Santos & Ramos, 2017), (SAS Institute, 2006). Uma forma habitual de avaliar um modelo é aplicá-lo a uma amostra dos dados, designada como dados de validação, sendo que, para um modelo ser considerado válido deverá funcionar para essa amostra de validação, bem como para os dados de treino utilizados na construção do modelo (SAS Institute, 2006).

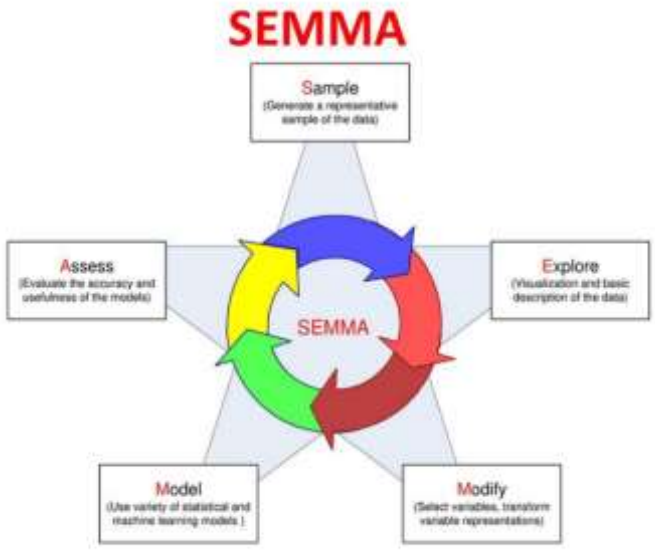

Figura 6 - Etapas constituintes do processo SEMMA

Fonte: Manohar, 2019[7]

[7] Consultado em http://www.mswamynathan.com/2019/11/01/accelerate-analysis-insight-data-mining/ a 04/04/2020

3.1.3 KDD

Por último, o **KDD** (*Knowledge Discovery in Databases*) remonta a 1989, que pode ser definido, por Fayyad, Shapiro & Smyth, 1996 como "um processo não trivial de identificação de novos padrões válidos, úteis e compreensíveis" (Camilo & Silva, 2009). Surge da necessidade de prestar auxílio à análise, interpretação e identificação de possíveis relacionamentos a partir de quantidades massivas de dados. Tal como o nome indica, foca-se sobretudo na descoberta e extração de conhecimento a partir dos mesmos, permitindo desenvolver e delinear estratégias de ação em função do contexto de aplicação (Goldschmidt & Passos, 2005), (Camilo & Silva, 2009), (Fávero, 2019).

O KDD é, ainda, considerado um processo interativo, e iterativo. É interativo no sentido, em que pressupõe a necessidade de intervenção humana no controlo de todo o processo e, iterativo, já que pode ser repetido parcial ou integralmente as vezes que se acharem necessárias, na ambição de atingir melhores resultados (Azevedo & Santos, 2008), (Vasconcellos, 2017), (Goldschmidt & Passos, 2005), (Camilo & Silva, 2009).

Existe uma clara distinção entre *Data Mining* e KDD que deve ser ressalvada, uma vez que, o KDD constitui todo o processo inerente à descoberta de conhecimento, enquanto que o *Data Mining* corresponde a uma das atividades desse processo (Goldschmidt & Passos, 2005), (Camilo & Silva, 2009), (Fávero, 2019). Como defendido por (Fayyad *et al.*, 1996), essa descoberta de conhecimento acontece recorrendo a técnicas de *Data Mining* e são necessários cinco passos para esse fim, encontrando-se abaixo brevemente explicado e ilustrados na **Figura 7**.

(1) Seleção de dados: etapa inicial que consiste em selecionar uma amostra dos dados considerados mais relevantes e imprescindíveis à análise que se pretende concretizar (Azevedo & Santos, 2008), (Vasconcellos, 2017), (Goldschmidt & Passos, 2005).

(2) **Pré-processamento**: fase de limpeza de dados e seleção de atributos relevantes para que estes apresentem a consistência, coerência e formato necessário. Engloba formas de correção de dados ausentes, redundâncias, lacunas nos dados ou eliminação de dados irrelevantes ou *outliers*, de forma a não comprometer a qualidade do processo (Azevedo & Santos, 2008), (Goldschmidt & Passos, 2005).

(3) **Transformação**: etapa onde ocorre uma análise aos dados obtidos na fase anterior e, dependendo ou não dessa necessidade, efetua-se uma reorganização ou transformação de dados, utilizando para isso diversos métodos, como a criação de novos atributos, integração de dados provenientes de outras fontes ou ainda redução de dimensionalidade dos mesmos, ao eliminar variáveis consideradas irrelevantes (Azevedo & Santos, 2008), (Fayyad *et al*.,1996).

(4) **Data Mining**: fase característica de definição de técnicas e algoritmos de análise e descoberta de dados a serem aplicados, em concordância com aquele que é o objetivo/tarefa de *Data Mining*. Visa descobrir *insights* como padrões, tendências, relações, entre outros, com o propósito de se extrair informação e conhecimento implícitos, relevantes e úteis (Azevedo & Santos, 2008), (Fávero, 2019), (Goldschmidt & Passos, 2005). Redes neuronais, algoritmos genéticos ou modelos estatísticos constituem alguns exemplos de técnicas que se podem aplicar nesta fase (Goldschmidt & Passos, 2005).

(5) **Interpretação/avaliação**: consiste numa fase de interpretação e avaliação dos *insights* anteriormente identificados, havendo a possibilidade de se regressar a uma das etapas anteriores. Esta interpretação converte-se e reflete-se em conhecimento contínuo, que se vai incorporando na organização, sustentando, e facilitando o processo de tomada de decisões (Fayyad *et al*.,1996).

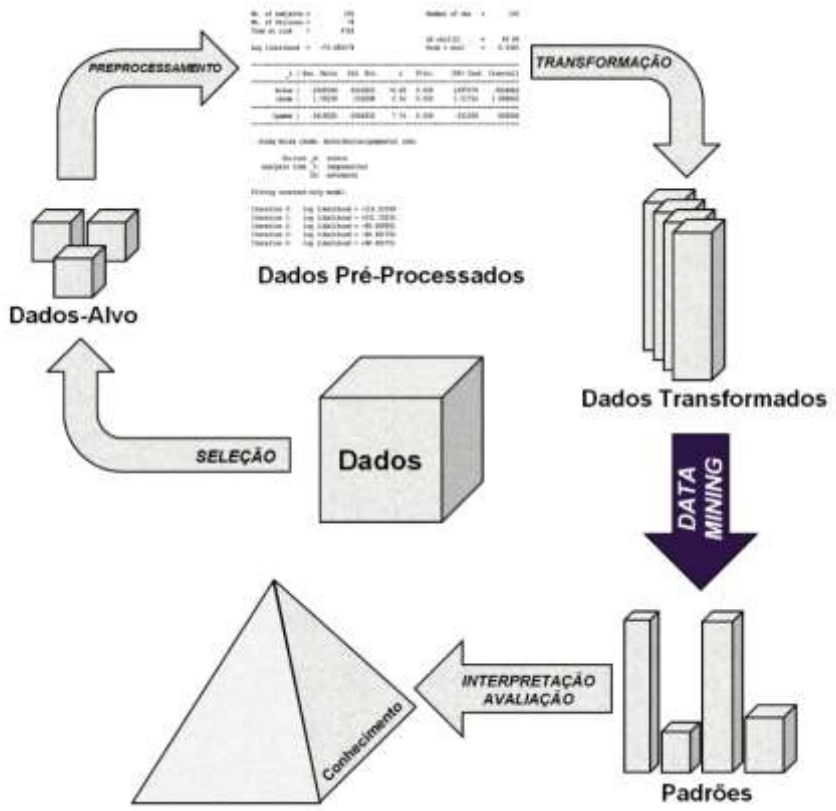

Figura 7- Estágios do processo KDD

Fonte: Fayyad, Piatetsky-Shapiro e Smyth, 1996

Apesar de serem três processos diferentes, cada um com as suas particularidades, no geral, apresentam essência e estrutura semelhantes (Goldschmidt & Passos, 2005).

3.2 Conceção de learning objects

Os *Learning Objects* (LO), ou, em português, objetos de aprendizagem, são uma estratégia de partilha de conhecimento para a exploração de tecnologia e de algoritmos, associados a uma área de alguma complexidade cognitiva. Existem vários modelos de criação de LO em função da estratégia pedagógica e do objetivo formativo.

Se a estratégia é informar de forma clara, como explorar um recurso, o modelo SOI **(S**eleção**, O**rganização**, I**ntegração – ***S****electing*, ***O****rganizing*, ***I****ntegrating***)** de

Mayer (1999) é especialmente vocacionado. De facto, o ambiente informativo e formativo inspirado neste modelo tem o objetivo de promover uma aprendizagem construtivista, (Reigeluth, 2009) e individual (Wannapipat, Chaijaroen & Somabut, 2013), onde, a construção de conhecimento é feita através de lições e instruções declarativas do mesmo, considerando o formando bastante autónomo e auto-construtor. Este material formativo é adequado para processos de aprendizagem baseados em texto, palestras e ambientes de natureza multimédia (Lima & Capitão, 2003), (Reigeluth, 2009). Como tal, servirá de base na conceção e elaboração dos LO, sob a forma de diversos tutoriais expostos em texto e auxiliados por vídeo presentes neste site, preparado para o público-alvo.

Para que tal ocorra, é necessário ativar três processos cognitivos fundamentais (Reigeluth, 2009):

(1) Selecionar a informação considerada relevante, ajudando os formandos a focarem no que é realmente essencial de reter e aprender, sendo claro, conciso e consistente (Lima & Capitão, 2003), (Reigeluth, 2009). A título de exemplo, a aplicação de diferente formatação à informação considerada chave, como a utilização de itálico, negrito, sublinhar, ou até mesmo a construção de um resumo simples, removendo o conteúdo supérfluo, constituem técnicas passíveis de serem aplicadas (Wannapipat *et al*., 2013).

(2) Organizar a informação para que o formando construa uma representação mental visual (pictórica) e verbal coerentes (Lima & Capitão, 2003), (Reigeluth, 2009), (Wannapipat *et al*., 2013). Estruturar o texto de forma breve, simples e precisa, fazendo uso de enumerações ou representações gráficas, multimédia ou animações são alguns dos métodos aplicados (Wannapipat *et al*., 2013).

(3) Integração da informação é o último processo, onde se pretende que o formando efetue uma representação integrada de todo o conteúdo e material aprendido, aliado ao conhecimento prévio de que dispunha

(Reigeluth, 2009), (Wannapipat *et al*., 2013), disponibilizando ao mesmo, exemplos, questões e desafios que potenciem e avaliem a sua interiorização (Lima & Capitão, 2003), (Reigeluth, 2009).

3.3 Design metodológico adotado

O *design* metodológico adotado para este projeto de investigação inspirado no *Design Science Research* consiste em síntese no teste ao *Power BI* das suas capacidades de análise preditiva que se encontram ao alcance de utilizadores não especialistas.

Assim sendo, à luz da abordagem e explicação dos processos supramencionados, o KDD é o que melhor se adequa e aplica na condução do objetivo desta dissertação, tendo em conta que o mesmo passa por extrair conhecimento útil, a partir da aplicação de algoritmos, de forma a melhorar a qualidade da tomada de decisões. O mesmo visa explorar um contexto mais elementar de acesso a uma base de dados experimental, sem processamento do lado do SQL e com uma versão do *Power BI* que não é *Premium*, com o objetivo de explorar o que se pode obter com recursos mais elementares. Apesar destas restrições, o cenário de trabalho não é, como se irá analisar, elementar, contudo, o KDD é considerado um guia adequado.

A explicitação da aplicação dos recursos *Power BI* será materializada tendo em conta as recomendações do Modelo SOI.

Capítulo 4
Ensaio do Power BI na análise preditiva

4. Exploração do Power BI Desktop

4.1 Enquadramento

No seguimento do trabalho elaborado, este capítulo destina-se exclusivamente à aplicação prática dos modelos anteriormente explicados, seguindo para isso, a apresentação de um tutorial elaborado, de forma a conseguir efetuar-se uma análise preditiva rigorosa e consistente, que sirva de base e suporte a uma consequente tomada de decisões mais seguras e sustentadas. Esta análise será abordada segundo um recurso principal e tradicional disponibilizado pelas funcionalidades incluídas no *PBI Desktop*, nomeadamente através do *line chart* contido no painel de análise. A apresentação e descrição dos recursos é acompanhada de uma reflexão crítica, que será retomada nas conclusões, razão pela qual, não se destaca uma secção de "discussão de resultados", dado também a natureza do estudo e do "produto" desenvolvido.

Para viabilizar o estudo, foi escolhida a BD *Adventure Works*, proveniente do *Microsoft Access*, que consiste num *Data Warehouse*, facultado pela *Microsoft* que contém dados de uma empresa fictícia cujo *core business* se prende com a venda de materiais desportivos de aventura, como bicicletas, roupas desportivas e acessórios, bem como a produção de algumas peças. É constituída por quatro áreas principais: vendas, produção, recursos humanos e compras/stock.

A par com o Cap.3, e tal como o **KDD** sugere no desenvolvimento dos testes de aplicação dos algoritmos de predição nas circunstâncias técnicas já descritas anteriormente, aplicaram-se as diferentes etapas:

(1) Seleção de dados: Usou-se a Base de dados *Adventure Works* dividida nas áreas de vendas, produção, recursos humanos e compras. O departamento que serviu de base a este estudo foi o de compras, mais especificamente o *PurchaseOrderDetail*, isto é, os detalhes dos pedidos de compra, nomeadamente, o *Purchase Order* (número do pedido), *Purchase Order Detail* (Número de item), *Due Date* (Data de remessa), *Order Quantity* (Quantidade), *ProductID* (código do produto), *UnitPrice* (preço unitário), *LineTotal* (Total da linha), *ReceivedQty* (Quantidade recebida), *RejectedQty* (Quantidade rejeitada), *StockedQty* (Quantidade em *stock*) e *ModifiedDate* (Data de modificação de pedido). É importante sublinhar que o âmbito da análise preditiva cobrirá a tabela *Purchasing_PurchaseOrderDetail*, mais concretamente o indicador *Line Total*.

(2) Pré-processamento: Após uma análise dos dados, verificaram-se algumas inconsistências na tabela em questão, que não cumprem os requisitos necessários a uma análise preditiva com qualidade, nomeadamente em relação aos anos de 2001, 2002, 2003, e 2005, sendo que:

- 2001 só apresentava dois meses;
- em 2002 estava em falta o mês de dezembro;
- em 2003 estavam em faltavam os meses de janeiro, fevereiro e abril;
- em 2005 existiam 57 linhas só com uma data e num formato diferente dos restantes;
- em 2004 existem todos os meses consecutivos até setembro e apesar de estarem três meses em falta não comprometem os resultados da análise em questão.

Posto isto, passou-se para a fase de limpeza de dados, onde, no ficheiro *Access* original se passou para a remoção dos anos de 2001 e 2005 por demonstrarem pouco rigor e solidez, impedindo a correta elaboração da previsão. Uma solução passaria por introduzir valores possíveis, já que se pretende analisar o potencial da tecnologia e não analisar um caso real e

sobre ele tirar conclusões. Tendo em conta que nestes anos o volume de compras foi muito baixo, assumiu-se como melhor prática a eliminação destes dados.

No ano de 2002 acrescentou-se uma linha que corresponde ao mês de dezembro em falta, mas com valor nulo apenas para se completar o ciclo de um ano.

Aplicou-se o mesmo princípio ao ano de 2003, acrescentando-se três linhas para os meses respetivos em falta apresentando sempre valor nulo, já que não existiu qualquer pedido nesses meses. Procedeu-se a essa prática, sublinhando uma vez mais com o intuito de satisfazer e reunir as exigências da análise-alvo.

Note-se que existe um leque vasto de soluções possíveis para colmatar a ausência de dados, dependendo do contexto e enquadramento empresarial, como substituição desses dados por valores médios, por valores mais prováveis ou mesmo por valores mais próximos, o que, neste caso específico poderia levar a um enviesamento dos resultados, já que os meses em falta são sinónimo de ausência de pedidos de compra.

(3) Transformação: Tendo a BD consistente, procedeu-se a algumas tarefas de transformação de dados aplicadas em PBI, nomeadamente:

- Na vista Dados, criou-se uma coluna nova denominada *Month*, através da aplicação da fórmula *"= FORMAT([DueDate], "YYYY-MM")"* em DAX para todas as linhas da BD. O *output* gerado reflete-se numa alteração de formato da coluna *DueDate*, de Dia-Mês-Ano para a forma Mês-Ano, agrupando assim o total de compras de cada mês (*Line Total*).
- Seguidamente, modificou-se o formato da coluna Month para Data/Hora, com o intuito de que passasse a ter validade cronológica, e sendo assim passível de aplicar no Eixo X;

- Por último, ordenou-se a mesma de forma ascendente, facilitando a visualização dos meses de forma ordenada.

(4) Data Mining: o principal objetivo/ problema de DM é a previsão, e como tal, serão utilizadas técnicas/métodos apropriados, como é o caso de modelos de previsão em séries temporais, isto é, fazer-se-á uso de modelos estatísticos, que no contexto específico, serão os *Exponential Smoothing models* (modelos de alisamento exponencial). Como já foi mencionado, neste tipo de modelos, a variável é prevista tendo em conta uma análise ao seu histórico, com o propósito de capturar algum tipo de padrão que possa vir a ser extrapolado na projeção do seu comportamento futuro. Para o fim, e de acordo com as componentes verificadas nos dados, recorrer-se-á à utilização de dois algoritmos distintos: **Seasonal algorithm** (**ETS AAA)**, e ou do **Non-Seasonal algorithm** (**ETS AAN)**.

(5) Interpretação/avaliação: esta fase destina-se à avaliação e interpretação dos *outputs* dos modelos de previsão gerados, tendo em conta aferir a qualidade e precisão dos mesmos. Será efetuada numa fase final, após comparação dos diversos modelos obtidos a partir da variação dos vários parâmetros de configuração.

Do mesmo modo, e como mencionado no capítulo anterior, a aplicação deste cenário de exploração de iniciação será levada a cabo através da criação de *LO*, sob a forma de um tutorial e vídeos demonstrativos. O mesmo pode ser encontrado no decorrer deste trabalho e, em paralelo, num *site*, propositadamente criado cujo *layout* pode ser consultado no apêndice C. Tem como base o modelo de SOI, sendo verificadas as subsequentes fases:

(1) Selecionar a informação considerada relevante, incluindo uma breve explicação teórica dos reputados aspetos chave da análise em questão, organizada em vários menus dispostos no *site*;

(2) **Organizar a informação** sob a forma de esquemas com diversas etapas de execução sucintamente enumeradas e acompanhadas por figuras e um vídeo demonstrativo;

(3) **Integração da informação** ao disponibilizar pequenos desafios propostos a serem resolvidos pelo aluno, de forma a testar o seu nível de aprendizagem e a eficácia deste método.

Sendo o PBI considerada uma ferramenta *self-service* e intuitiva, note-se que, esta demonstração prática fará uso da versão mais elementar da mesma, acabando por não cobrir totalmente a área mais sofisticada da ferramenta *Premium*, à qual não se teve acesso, nem a um contexto de acesso a dados tecnicamente sofisticado, com "*views* de SQL" podendo indubitavelmente comprometer a qualidade dos seus resultados e a o seu potencial com um leque mais rico e diversificado de soluções. O principal objetivo é o de, dentro de um cenário de análise preditiva, identificar e aferir a capacidade e as funcionalidades que a mesma disponibiliza e estão ao alcance de utilizadores iniciantes nesta área, nomeadamente os que integram o grupo *End Users* e, por isso, não dispõem de conhecimento profundo a nível estatístico, matemático ou de linguagem SQL nem de meios tecnológicos muito sofisticados e devidamente integrados. Todo este enquadramento está representado no apêndice C.

4.2 Análise preditiva com recurso ao line chart

Como referido, a análise preditiva será efetuada com o recurso proveniente e integrado no *Power BI*, através do gráfico de linhas ou *line chart*, disponível no painel de análise.

Desta forma, será selecionado um indicador considerado interessante de prever no contexto em concreto, com base nos seus dados históricos e criados diferentes modelos com variação de parâmetros de configuração, que serão

posteriormente analisados e interpretados para, numa última instância, aferir o contributo e relevância dos mesmos para a análise preditiva. Antes de se partir para a elaboração do tutorial de forma prática, é fundamental que se compreenda como tudo é concretizado na ferramenta.

Infelizmente, e até à data, desconhece-se ao certo o algoritmo exato que as características de previsão do *Power BI Desktop* utiliza, sendo considerado como uma *black box*, isto é, um processo invisível que não tem uma explicação clara quanto ao seu funcionamento. Contudo, existem várias explicações pormenorizadas acerca do algoritmo que o *Power View* utiliza neste tipo de análise, sendo essa a explicação que se irá adotar e assumir no decorrer desta demonstração, já que, são diversos os autores defensores de que, provavelmente, a ferramenta de previsão do *Power BI* utiliza o mesmo princípio aplicado em *Power View* (Burki, 2018), (Smoak, 2019).

O *Power View* disponibiliza aos seus utilizadores ferramentas de análises estatísticas avançadas, como é o caso da previsão (FluentPro, 2019). A previsão aplicada a séries temporais em *Power View* baseia-se e faz uso de um conjunto de modelos de previsão incorporados e provados como eficazes: *Exponential Smoothing Models*, anteriormente explicados. Estes modelos são mundialmente utilizados em diversos domínios, objetivam detetar padrões de sazonalidade de forma automática bem como captar tendências de forma eficiente com o propósito de os extrapolar para o futuro. Ajudam ainda, a colmatar valores desviados (*outliers*), e a identificar irregularidades ou imprevisibilidades contidas nos dados, sendo estas últimas suavizadas, suprimidas e eliminadas pelo próprio modelo, com o fim de não comprometer ou distorcer o mesmo e os seus *outputs* de previsão gerados para o conjunto de dados em análise (Microsoft, 2014), (FluentPro, 2019), (Smoak, 2019), (Powell, 2017).

Associados aos modelos de alisamento exponencial, existem duas versões de algoritmos implementados de acordo com as características dos dados, sendo estes o **Seasonal algorithm** (**ETS AAA)** aplicado a dados que além de tendência,

apresentem comportamentos de sazonalidade, comummente chamado algoritmo **Holt-Winters**. Faz uso de uma equação onde combina as componentes de erro, tendência e sazonalidade de forma aditiva. E o **Non-Seasonal algorithm** (**ETS AAN)**, aplicado a dados que reflitam apenas comportamentos de tendência, isentos de qualquer sazonalidade (Burki, 2018), (Microsoft, 2014), (FluentPro, 2019). Este algoritmo faz uso de uma equação mais simples, onde, combina apenas as componentes de erro e tendência de forma aditiva, desconsiderando a sazonalidade (Microsoft, 2014), (FluentPro, 2019). Por conseguinte, o *Power View* seleciona e recorre ao modelo que melhor se adequa de forma automática, após uma análise aos dados históricos (Microsoft, 2014), (FluentPro, 2019).

Existem quatro parâmetros ou *inputs* principais que visam otimizar a *performance* dos modelos e algoritmos, e, para tal, é crucial que a sua configuração seja feita de forma assertiva, para que, o PBI consiga extrapolar corretamente os *outputs* que daí advêm (FluentPro, 2019). Sendo estes:

a) **Forecast Length (Horizonte temporal de previsão) (FL/HT):** refere-se ao alcance da previsão, e em PBI, por defeito, ao número de pontos que se objetivam prever. Claro que vão depender da unidade de tempo que se está a utilizar, isto é, se a análise estiver a ser feita mensalmente, então, o alcance de previsão será configurado em meses, aplicando-se o mesmo princípio a diferentes unidades de tempo (FluentPro, 2019).
b) **Ignore Last/Hindcast (Ignorar último) (IL/IU):** permite que se excluam determinados períodos, sendo um parâmetro bastante útil. Por um lado, torna-se extremamente vantajoso quando se ignoram períodos passados, já que, quando efetuado, o PBI vai incluir na sua previsão períodos com valores já conhecidos que em termos práticos se traduz numa forma de testar a precisão das previsões, ao prever valores passados (Powell, 2018). Estes valores, são, numa fase posterior comparados com os valores observados/reais e daí se avalia o desempenho do modelo. Ao ignorar o último, o algoritmo só vai ter em

conta os dados até aos períodos que foram ignorados, excluindo-os (Powell, 2018), (Mehta, 2017). Assim, quantos mais períodos se ignorarem, menos informação é disponibilizada e menor será a representação da previsão (Microsoft, 2014). É utilizado ainda para ignorar períodos incompletos ou errados (Powell, 2018), (FluentPro, 2019). Se, eventualmente, os valores reais para os períodos que foram previamente ignorados não se encontrarem dentro do IC para os valores de previsão, a previsão pode não ser válida ou deverá alterar-se o IC para um nível maior (Powell, 2018).

c) **Confidence Interval (Intervalo de Confiança (IC)):** esta opção permite definir um valor de probabilidade estatística que determina a probabilidade de um valor real estar próximo do valor de previsão. Ou seja, se o IC for de 95%, isto significa que existem 95% de hipóteses de o valor real estar dentro do intervalo de valores previstos, cujo, por norma, se define como sendo uma região sombreada. Quanto maior é o IC, menor é a margem de erro (FluentPro, 2019). Consequentemente, maior será a área abrangida pelo IC e, assim, maior será o *Upper Bond* e menor o *Lower Bound* (Powell, 2018). O IC está diretamente ligado ao *Upper* e *Lower Bond*, sendo que, o *Upper Bond* corresponde ao limite superior, e por isso ao valor máximo que o valor de previsão pode assumir e, o *Lower Bond* corresponde ao limite inferior e por isso, ao valor mínimo que o valor de previsão poderá assumir, constituindo desta forma o IC para os valores de previsão (FluentPro, 2019).

d) **Seasonality (Sazonalidade):** como referido anteriormente, a sazonalidade corresponde a fenómenos que ocorrem quando os dados exibem um tipo de comportamento específico que se repete de forma semelhante a cada período idêntico, na forma de ciclos, correspondendo estes normalmente a um ano. A sazonalidade pode ser aferida através de um ciclo de dados, correspondendo ao número de

pontos do referido ciclo. Exemplificando: se um utilizador estiver a analisar dados de vendas que flutuem ao longo de um ano, mas que tendam a comportar-se de forma semelhante, num dado período, ano após ano, então essa série temporal apresenta uma sazonalidade de um ano. Esta opção permite especificar a sazonalidade de forma manual, se o analista tiver conhecimento suficiente para disponibilizar essa informação ao algoritmo. Assim, para um ciclo de um ano, a sazonalidade seria de 365 dias, no caso da unidade de tempo estiver estabelecida em dias, de 52 semanas se a unidade de tempo estiver estabelecida em semanas, e assim sucessivamente, estando a sazonalidade dependente do número de unidades que constituem um ciclo, isto é, da sua duração, bem como da unidade de tempo utilizada (FluentPro, 2019), (Microsoft, 2014), (Powell, 2017). Ao especificar um valor para a sazonalidade, esse número é incorporado na fórmula de cálculo de *insights*, aumentando a capacidade que a previsão demonstra em compensar valores *outliers*. A sua estimação tem um impacto bastante elevado em previsões de séries temporais, especialmente em algoritmos utilizados em modelos de alisamento exponencial, uma vez que estes, requerem a sazonalidade como um *input* e se mostram habitualmente sensíveis a variações destes valores (FluentPro, 2019), (Microsoft, 2014). Uma vez conhecido o seu valor, é recomendável que se aplique manualmente para potenciar a precisão do modelo (Powell, 2017).

4.3 Aplicação prática em PBI Desktop

Tendo já sido feita uma contextualização da BD e breve explicação do funcionamento em PBI, este subcapítulo destina-se à reprodução prática em *Power BI* da explicação concetual dos subcapítulos anteriores.

4.3.1 Tutorial

Em PBI *Desktop*, para se efetuar a análise preditiva com recurso ao *line chart*, é necessário que se verifiquem três requisitos principais antes de todo o processo, nomeadamente:

- O valor do eixo dos x deverá estar em formato data/tempo (nunca podendo estar em formato texto) ou disponibilizar quaisquer outros números inteiros, onde, em ambos os contextos, os pontos deverão ser equidistantes entre si (Powell, 2017), (Microsoft, 2014), (Guilfoyle, 2017);
- É necessário ter um mínimo de seis pontos de dados (coordenadas) (Powell, 2017), (Guilfoyle, 2017);
- A ferramenta de previsão apenas se encontra disponível para o gráfico de linhas (*Line Chart)* e funciona para um único indicador/linha (Powell, 2017), (Microsoft, 2014), (Guilfoyle, 2017).

- **Análise preditiva face ao total de compras (*line total*) em euros, para um horizonte temporal de 12 meses, ignorando os últimos seis.**

O indicador volume total de compras (*Line total)* proveniente da tabela *Purchasing_PurchaseOrderDetail* permite verificar o valor gasto em compras da atividade da empresa (exclui o valor dos custos de transporte e impostos). Este indicador corresponde sempre ao somatório do valor dos períodos em análise

(pode ser por ano, trimestre, mês, semana, dia). Neste caso, o objeto desta análise, será um indicador mensal e como tal criou-se a coluna *Month*, previamente explicada.

Note-se que, os parâmetros utilizados neste modelo são apenas como forma de exemplificação para a execução do tutorial, podendo os mesmos assumirem valores diferentes, de acordo com o objetivo de previsão, como vai ser efetuado no seguimento na construção de diversos modelos.

O vídeo demonstrativo que atesta o que seguidamente será explicado pode ser consultado no *site*.

1. O primeiro passo é o da conexão com a fonte de dados, ou seja, é nesta fase que ocorre a importação dos dados em que se vai trabalhar para o PBI. Tal, faz-se através do Menu "*Get data*" ou em português, "Obter dados" que se encontra na *interface* do *Power BI* (Mehta, 2017), na **Figura 8**:

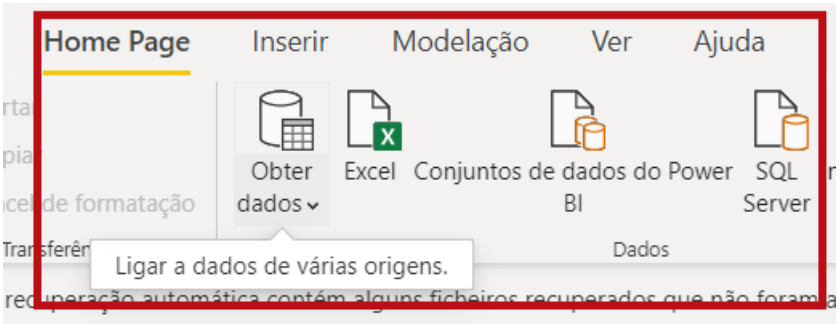

Figura 8- Conexão à fonte de dados

Ao clicar neste menu, ficam disponíveis diferentes fontes de dados possíveis, escolhendo o utilizador a fonte adequada aos seus dados.

2. Esta etapa é onde se deve assegurar que o primeiro requisito necessário à previsão está a ser satisfeito: depois de selecionada a variável que se quer

prever, neste caso, o *Line Total* e a variável temporal *Month*, em formato relatório, recorrer à visualização na forma de tabela e confirmar se os valores do eixo do x (variável *Month*) estão no formato requerido, ao clicar nessa mesma variável e olhando para o seu formato, como ilustra a **Figura 9** (Powell, 2017), (Guilfoyle, 2017):

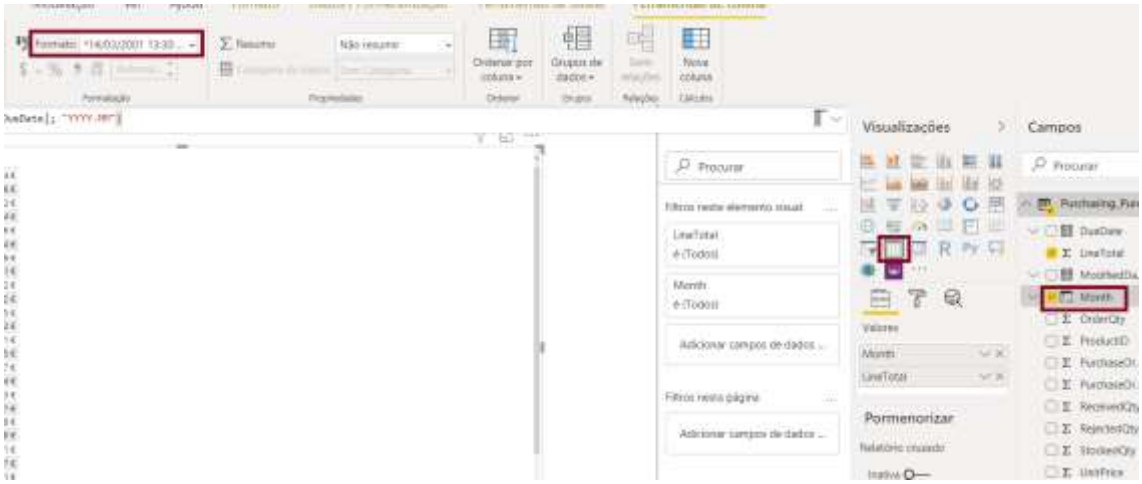

Figura 9-Confirmação do formato das variáveis

3. Estando a variável do eixo do x no formato requerido, escolher o gráfico *line chart* ou gráfico de linhas que se encontra nas visualizações, na vista de relatório (Powell, 2017), (Mehta, 2017), (Smoak, 2019), (Guilfoyle, 2017). A **Figura 10** indica o ícone que respeita à visualização requerida:

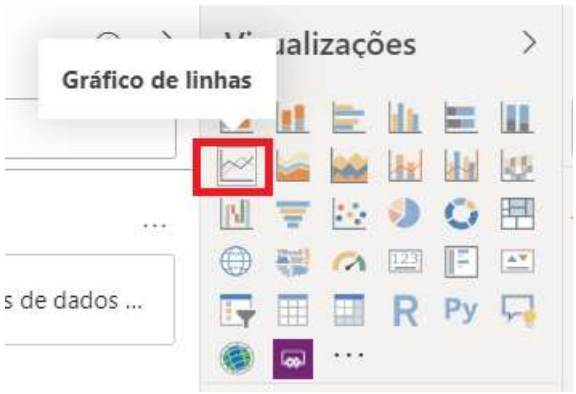

Figura 10- *Line chart* ou gráfico de linhas

Este tipo de gráficos só funciona para um indicador (para uma única linha) e é constituído por duas variáveis (Microsoft, 2014):

- o **eixo do y** (***value)***, que representa a variável numérica a escolher dos campos possíveis e, que se vai arrastar para este campo. De uma forma simplista, o eixo dos y representa o indicador que se quer prever, que no caso será o *Line Total* (Powell, 2017), (Guilfoyle, 2017);
- o **eixo do x** (***axis***) representa sempre a variável temporal, que, neste caso específico será a variável *Month*, que já apresenta uma estrutura mensal. Por defeito, o PBI cria uma hierarquia de calendário para o eixo do x com colunas para ano, trimestre, mês e dia, como se pode verificar na **Figura 11** (Powell, 2017), (Guilfoyle, 2017):

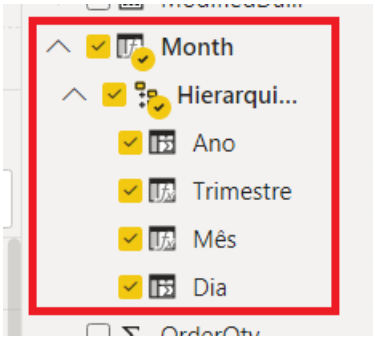

Figura 11- Hierarquia de calendário

Esta hierarquia poderá ser experimentada através dos botões existentes na parte superior do gráfico, clicando em **expandir tudo para um nível na hierarquia,** como indica a **Figura 12** (Powell, 2017):

Figura 12- Opção expandir tudo

Contudo, neste caso, escolheu-se apenas a opção mensal, de forma a que os dados se visualizem na forma mensal.

É através da visualização do gráfico de linhas, que se consegue confirmar outro dos requisitos relacionado com a existência de pelo menos seis pontos, neste caso num formato de data, equidistantes. Com o gráfico de linhas vazio e selecionado, arrastar as variáveis para os devidos campos (Powell, 2017).

4. Com o gráfico selecionado, abrir a **lupa de análise (*Analytics pane*)** que se encontra no painel de visualizações ao lado da formatação como indica a **Figura 13** (Powell, 2017), (Mehta, 2017), (Burki, 2018), (Smoak, 2019), (Guilfoyle, 2017):

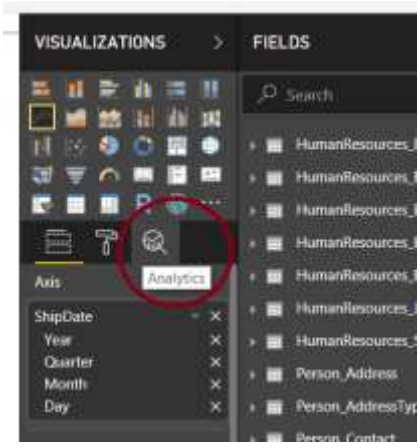

Figura 13- Lupa de análise

E, seguidamente, selecionar a linha de previsão (*forecast*) e clicar em **adicionar,** como ilustra a **Figura *14*** (Powell, 2017), (Mehta, 2017), (Burki, 2018), (Smoak, 2019), (Guilfoyle, 2017):

Figura 14- Linha de previsão

Por defeito, o *Power BI* efetua uma previsão para 10 pontos, ou seja, neste caso para 10 meses e é criada com um intervalo de confiança de 95%.

É nesta etapa que é possível definir os diversos parâmetros ou opções de configuração, nomeadamente, o alcance da previsão, a unidade de tempo a utilizar, assim como os períodos que se pretendem ignorar, o nível de confiança e a sazonalidade. Valores estes determinados de forma devidamente adequada ao contexto do utilizador (Powell, 2017), (Mehta, 2017), (Smoak, 2019), (Guilfoyle, 2017). No caso, alterou-se o **alcance da previsão** para 12 meses. O segundo parâmetro, como explicado anteriormente, possibilita que se **ignorem períodos**. Assim, ignoraram-se os últimos seis meses, o que significa que apenas se estão a disponibilizar dados ao algoritmo de previsão até esses seis meses, excluindo-os. Isto significa que, esses dados disponibilizados ao algoritmo de previsão, são utilizados para treinar o modelo, e, baseado nos mesmos, projetar valores previstos (Mehta, 2017). No que respeita ao **IC**, utilizou-se um nível de confiança de 95%, significando isto, que, o algoritmo, projetou os valores de previsão com 95% de hipótese do valor real se encontrar dentro desse intervalo, estando este representado pela área cinzenta do gráfico (Mehta, 2017), (Burki, 2018). Por

último, assumiu-se uma sazonalidade de 12 meses, visto que a unidade de tempo é o mês e se definiu um ano como sendo a duração de um ciclo. Estes critérios encontram-se definidos na **Figura 15**. Tendo todos critérios preenchidos, basta **aplicar**:

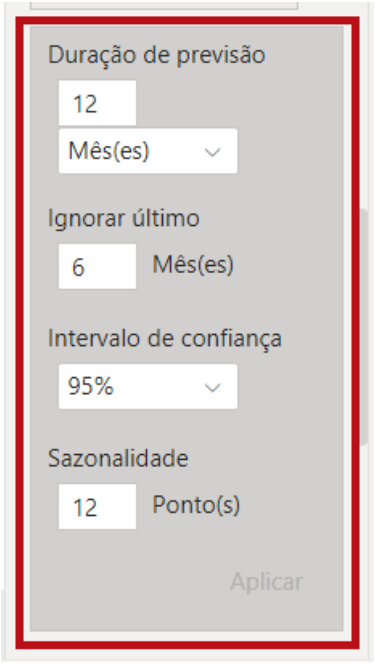

Figura 15 - Parâmetros de configuração

5. Por último, existe a possibilidade de fazer uso das opções de formatação para dar vida aos gráficos alterando características como cor, estilo, tamanho de letra, transparência e banda de confiança de forma a facilitar a distinção entre valores reais e valores de previsão, realçando estes últimos (Powell, 2017).

6. A última fase, para efeitos de análise, consiste em exportar os dados obtidos em PBI. Basta clicar no Menu de mais opções, representado pelas eplipses (reticências), que se podem encontrar abaixo ou acima do gráfico, seguidamente clicar em exportar dados, como exemplifica a **Figura 16** , que, posteriormente guardará toda a informação num ficheiro CSV que poderá ser consultada.

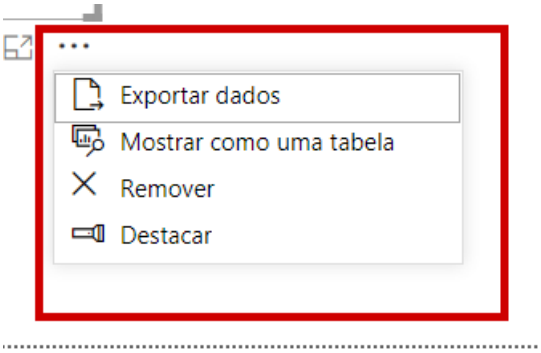

Figura 16- Exportar dados

Posto isto, o *output* conseguido neste e noutros modelos será seguidamente analisado, interpretado e avaliado. Todos os modelos seguiram os mesmos passos, alterando apenas os parâmetros de configuração.

4.3.2 Interpretação e avaliação

No seguimento do processo KDD, este subcapítulo constitui a última fase, e por isso, a fase de interpretação e avaliação dos modelos e respetivos *outputs*. Como já foi referido anteriormente, todos os parâmetros de configuração apresentam um impacto e relevância significativa no contexto da análise preditiva. Contudo, existem dois indicadores muito competentes na avaliação da qualidade e precisão das previsões: o *hindcast* e a variação dos IC (Microsoft, 2014). O *hindcast* é utilizado para avaliar o desempenho do algoritmo no contexto de previsão, mostrando como a previsão teria previsto o passado recente a partir do momento escolhido. É através do mesmo que se consegue efetuar uma comparação entre o valor previsto e o valor real, e, perceber se o primeiro se distancia muito do segundo. É de ressalvar que, previsões com boa qualidade não são aquelas que coincidem exatamente com os valores reais, uma vez que, quando um modelo reflete valores e tendências exatamente coincidentes com os valores reais, este pode estar manipulado/*overfitted* e como tal, as suas projeções

ou prospeções podem não constituir um bom indicador de qualidade da previsão. Por outro lado, num cenário em que se verifiquem valores previstos suficientemente semelhantes e próximos aos reais, sem coincidirem exatamente com os mesmos, constitui um melhor indicador de qualidade da previsão (Microsoft, 2014). O segundo indicador ilustra, de forma visual, a fiabilidade de uma previsão (Microsoft, 2014). Assim, o seu objetivo consiste em fazer variar esta probabilidade, experimentando diferentes níveis de confiança, na tentativa de se compreender o impacto esperado nos resultados de previsão (Microsoft, 2014). A variação do nível de confiança, terá impacto na área a sombreado que, por si só, representa o intervalo para os valores de previsão, sendo esta maior à medida que se aumenta o nível de confiança (Powell, 2018).

Tendo em conta a explicação supramencionada e o objetivo desta análise, foram gerados e testados diversos modelos, fazendo variar os vários parâmetros de configuração, com o propósito de aferir o impacto e influência que os mesmos poderão vir a ter neste tipo de análise.

4.3.2.1 Teste ao parâmetro ignorar último (IU)

Numa primeira instância, procedeu-se a um teste ao parâmetro ignorar último (IU), o que significa que se pretende aferir o impacto e importância que a variação deste parâmetro tem na análise preditiva. Assumiu-se que os parâmeros restantes se mantêm constantes. Posto isto, confrontaram-se os diversos modelos gerados ilustrados de seguida, onde, o HT representa o horizonte temporal da previsão, o IL o ignorar último, o IC o intervalo de confiança e a "saz." a sazonalidade.

Modelo 1)	**Parâmetros:**
	HT= 12 meses
	IL= 0
	IC= 95%
	Saz.= 12 meses

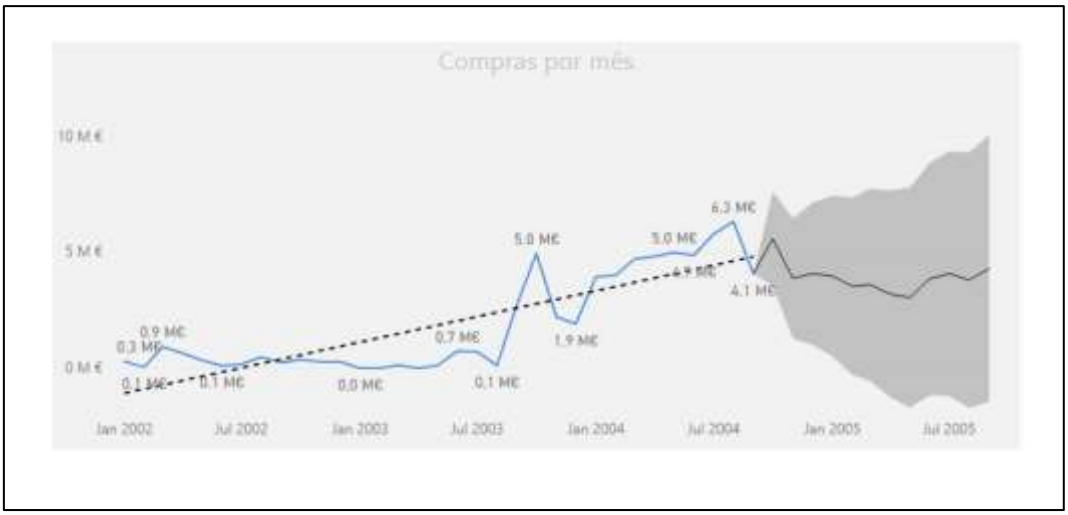

Figura 17- Modelo 1 do teste ao parâmetro IU

A partir da **Figura 17**, é possível verificar que, partindo dos *outputs* provenientes deste modelo, o *line total*, representado pela linha azul, corresponde aos valores observados/reais do indicador que se pretende prever. O total de compras, representado pela linha preta, corresponde aos valores previstos, onde, para cada ponto previsto, o *upper bond* representa o valor máximo que o valor previsto poderá atingir e o *lower bound* o valor mínimo que o valor previsto poderá atingir, sendo que estes valores constituem o intervalo para os valores previstos, representado pela área a sombreado.

A partir do gráfico acima, depreende-se que, para os meses de previsão existe uma reprodução aproximada do volume de compras correspondente ao período homólogo do ano anterior, revelando-se, contudo, mais estável e por isso com menos oscilações e com tendência crescente face ao período observado. Adicionalmente, observa-se a reprodução do pico ocorrido no mês 10 (outubro).

Modelo 2)	**Parâmetros:**
	HT= 12 meses
	IL= 6 meses
	IC= 95%
	Saz.= 12 meses

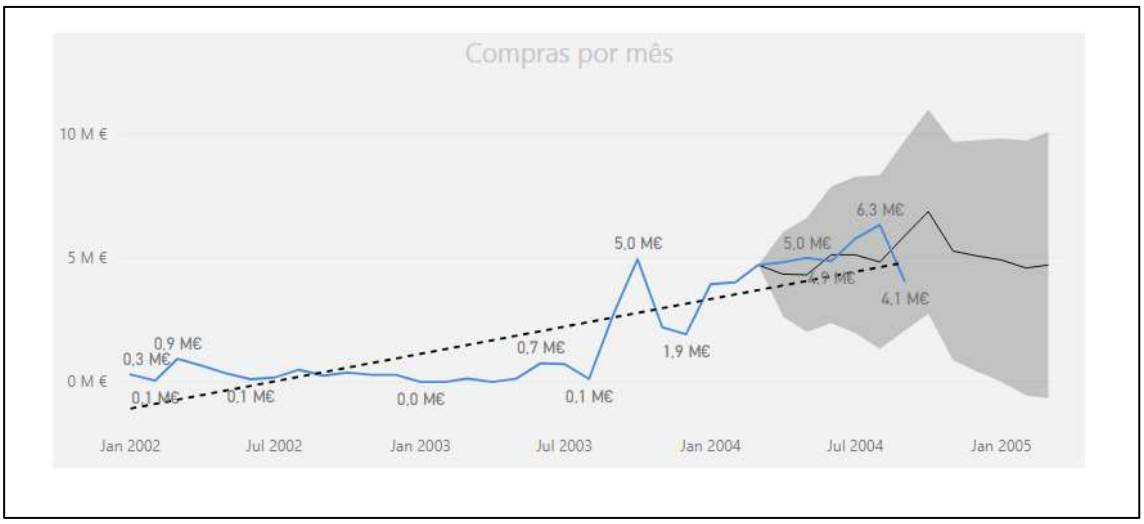

Figura 18- Modelo 2 do teste do parâmetro IU

No caso do modelo 2, representado na **Figura 18**, e tendo em conta que são ignorados seis meses, o algoritmo de previsão que gera este modelo, apenas irá considerar e utilizar como *input* os meses que antecedem o período ignorado. Assim, a previsão será de 12 meses, onde, os seis primeiros são passíveis de comparação com os valores observados. Atentando ao gráfico correspondente ao modelo dois e relativamente aos valores de previsão, observa-se uma reprodução semelhante aos meses de março a setembro do período homólogo ainda que estes sejam mais elevados, denotando uma tendência crescente. À semelhança do modelo anterior, verifica-se uma reprodução do pico ocorrido no mês 10 (outubro).

Como se ilustra no **Quadro 2** , o erro evolui de forma inconstante, concluindo-se que, em certos meses, os valores de previsão se aproximam dos valores

observados e, noutros, se distanciam, sendo a média do erro absoluto de aproximadamente 15%.

Modelo 3)	**Parâmetros:**
	HT= 12 meses
	IL= 8 meses
	IC= 95%
	Saz.= 12 meses

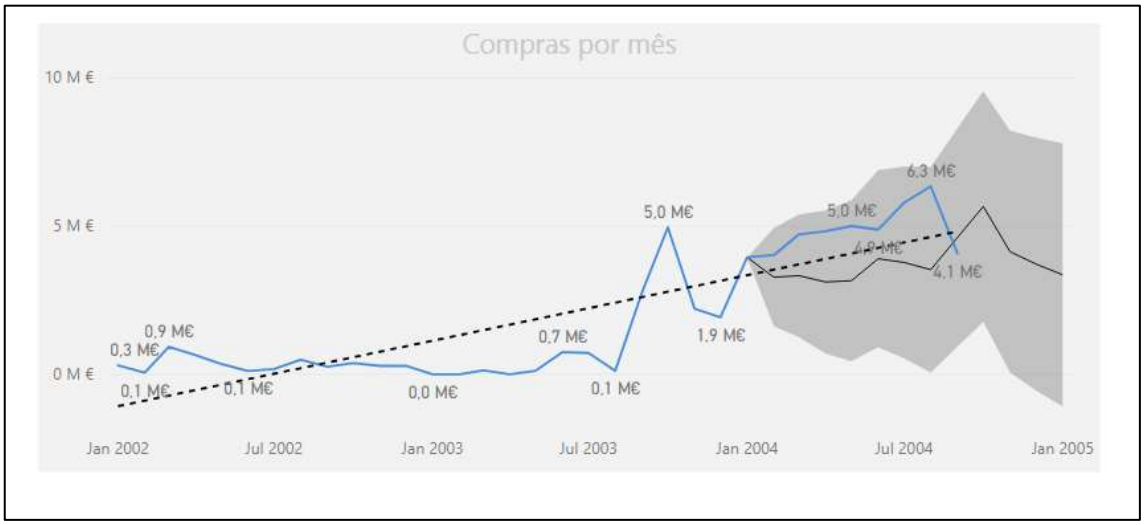

Figura 19 - Modelo 3 do teste ao parâmetro IU

No modelo três, ilustrado na **Figura 19**, são ignorados oito meses, e por isso, uma vez mais, o algoritmo para efetuar a previsão, irá basear-se na informação até ao período ignorado. Assim, a previsão será de 12 meses, onde, os oito primeiros são passíveis de comparação com os valores observados. Atentando ao gráfico acima, e relativamente aos valores de previsão, observa-se uma reprodução de comportamento semelhante ao período real, contudo, de valores mais baixos. Tal acontece uma vez que, e como supramencionado, o algoritmo apenas inclui os dados até ao período ignorado, isto é, até janeiro de 2004, e por isso, a previsão efetuada baseou-se em valores históricos mais baixos e tal reflete-se na projeção do seu comportamento futuro.

Nos meses seguintes, os valores de previsão apresentam um comportamento semelhante ao ano anterior, verificando-se uma reprodução do pico ocorrido no mês 10 (outubro) de valor muito próximo ainda que um pouco mais elevado, bem como dos meses seguintes onde se ilustra um claro decréscimo ainda que não tão acentuado.

Em conformidade com o verificado no modelo dois e a partir do ***Quadro 2***, o erro evolui de forma inconstante, sendo o distanciamento do valor de previsão face ao valor real, no geral, superior, como se ilustra através do valor médio do erro absoluto de aproximadamente 26%.

Modelo 4)	**Parâmetros:**
	HT= 12 meses
	IL= 12 meses
	IC= 95%
	Saz.= 12 meses

Figura 20 – Modelo 4 do teste ao parâmetro IU

A **Figura 20** ilustra o modelo quatro, onde são ignorados 12 meses, e por isso, uma vez mais, o algoritmo para efetuar a previsão, irá basear-se na informação até ao período ignorado, ou seja, até setembro de 2003. Assim, a previsão será de 12 meses, onde, todos são passíveis de comparação com os valores observados.

Atentando ao gráfico acima, e relativamente aos valores de previsão, observa-se uma reprodução de comportamento semelhante ao período real até onde o

algoritmo tem conhecimento. Consequentemente, o pico representado nos modelos anteriores não se verifica, uma vez que o algoritmo não recebeu dados provenientes a esses meses.

Desta forma, o erro evolui de forma inconstante, sendo o seu valor bastante elevado, como se atesta através do valor médio do erro absoluto de aproximadamente 79%, como se pode verificar no **Quadro 2** abaixo.

Quadro de Erros

Month	LineTotal	Modelo 2	Modelo 3	Modelo 4
setembro 03	2 756 843,22 €			0%
outubro 03	4 956 873,26 €			88%
novembro 03	2 211 744,23 €			77%
dezembro 03	1 925 623,83 €			72%
janeiro 04	3 944 008,85 €		0%	94%
fevereiro 04	4 021 179,91 €		19%	94%
março 04	4 720 815,55 €	0%	30%	91%
abril 04	4 829 574,58 €	10%	36%	94%
maio 04	5 002 330,47 €	14%	37%	93%
junho 04	4 880 124,06 €	5%	20%	86%
julho 04	5 786 081,74 €	11%	35%	88%
agosto 04	6 344 902,07 €	24%	44%	92%
setembro 04	4 077 395,67 €	44%	13%	58%
MAPE		**15%**	**26%**	**79%**

Quadro 2 – Quadro de erros

Comparando, graficamente, os diferentes modelos, é possível de verificar que os meses de previsão reproduzem o comportamento dos meses anteriores conhecidos, apesar dos valores esperados terem tendência para aumentarem, tal como a linha de tendência a tracejado o indica.

Através dos mesmos, comprova-se que, quantos mais períodos se ignorarem, menos informação e conhecimento é disponibilizado ao algoritmo, fazendo com que este efetue uma previsão baseada em menos dados, aumentando assim a

margem de erro e consequentemente o distanciamento entre os valores de previsão e os valores reais, comprometendo a qualidade de previsão.

4.3.2.2 Teste ao parâmetro intervalo de confiança (IC)

Procedeu-se a um teste ao parâmetro intervalo de confiança (IC), com o intuito de se aferir o impacto e importância que a variação deste parâmetro tem na análise preditiva. Assumiu-se que os parâmeros restantes se mantêm constantes. Posto isto, confrontaram-se os diversos modelos gerados, onde o HT, o IL, o IC e o Saz. apresentam o mesmo significado dos modelos anteriores.

Modelo 1)	**Parâmetros:**
	HT= 12 meses
	IL= 0
	IC= 95%
	Saz.= 12 meses

Figura 21 - Modelo 1 do teste ao parâmetro IC

Modelo 2) Parâmetros:

HT= 12 meses
IL= 0
IC= 99%
Saz.= 12 meses

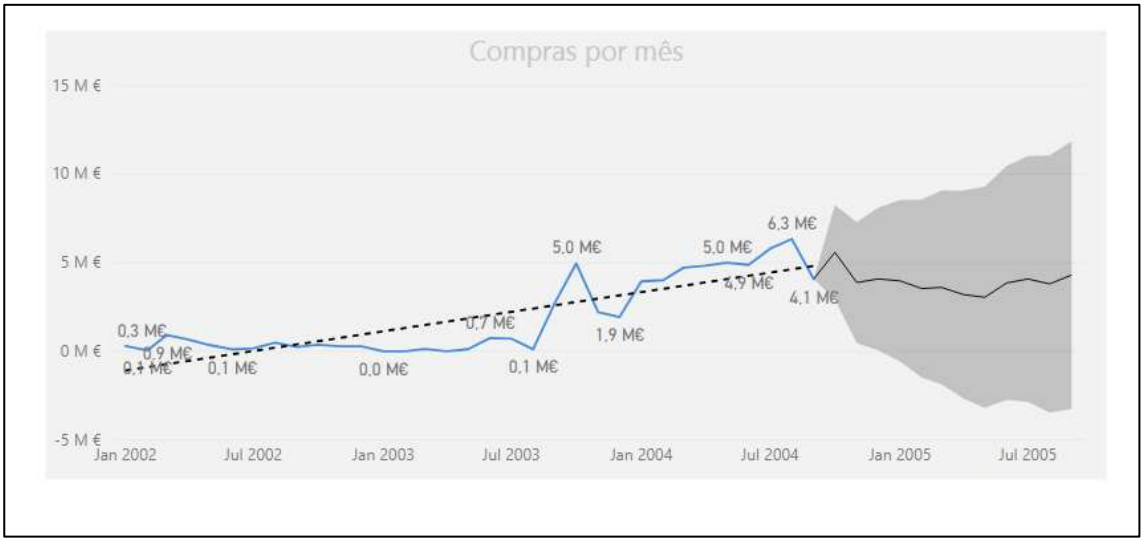

Figura 22 - Modelo 2 do teste ao parâmetro IC

Modelo 3) Parâmetros:

HT= 12 meses
IL= 0
IC= 90%
Saz.= 12 meses

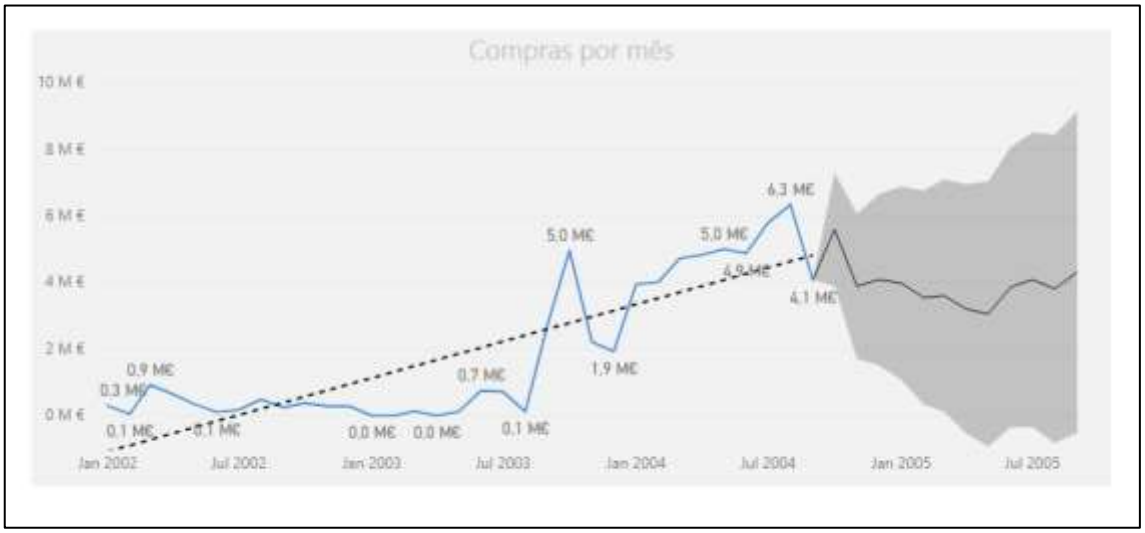

Figura 23 - Modelo 3 do teste ao parâmetro IC

Modelo 4) Parâmetros:

HT= 12 meses
IL= 0
IC= 75%
Saz.= 12 meses

Figura 24 - Modelo 4 do teste ao parâmetro IC

Data	Upper Bound				Lower Bound			
	IC 95	IC 99	IC 90	IC75	IC 95	IC 99	IC 90	IC75
set/04	4077395,67	4077395,67	4077395,67	4077395,67	4077395,67	4077395,67	4077395,67	4077395,67
out/04	7621026,77	8258603,94	7294719,16	6782480,80	3561856,39	2924279,22	3888164,00	4400402,36
nov/04	6482572,66	7299069,82	6064695,02	5408709,85	1284298,21	467801,05	1702175,84	2358161,02
dez/04	7150246,65	8112967,07	6657532,97	5884070,00	1021033,79	58313,38	1513747,48	2287210,44
jan/05	7447990,70	8537483,05	6890396,01	6015082,72	511677,32	-577815,03	1069272,01	1944585,30
fev/05	7374768,00	8577746,20	6759092,03	5792602,63	-284059,35	-1487037,55	331616,62	1298106,02
mar/05	7767897,93	9074542,30	7099166,33	6049390,07	-550925,87	-1857570,25	117805,73	1167581,98
abr/05	7668694,18	9071363,96	6950817,44	5823893,05	-1261480,64	-2664150,42	-543603,91	583320,49
mai/05	7801977,40	9294507,30	7038110,88	5838991,62	-1700296,82	-3192826,71	-936430,29	262688,97
jun/05	8878812,38	10456091,19	8071571,94	6804364,21	-1163020,55	-2740299,36	-355780,11	911427,62
jul/05	9361777,60	11019478,24	8513377,82	7181558,08	-1192065,38	-2849766,02	-343665,60	988154,15
ago/05	9328407,84	11062805,26	8440755,17	7047316,18	-1713729,39	-3448126,81	-826076,73	567362,26
set/05	10063024,14	11870867,44	9137782,39	7685335,95	-1446710,33	-3254553,62	-521468,57	930977,86

Quadro 3 – *Upper bound* e *lower bound* do IC

Comparando os diversos modelos ilustrados nas **Figura 21**, **Figura 22**, **Figura 23** e **Figura 24**, afere-se que os valores de previsão irão manter-se inalterados ao longo dos mesmos, independentemente do nível de confiança inerente, tendo em conta que, com a alteração do nível de confiança, apenas se altera a escala do

gráfico. Os mesmos, reproduzem o comportamento histórico a partir do mês de outubro de 2003, apesar de, no geral, assumirem valores mais elevados em relação aos valores observados.

Analisando o **Quadro 3**, acima, afere-se que, quando se aumenta o nível de confiança, no caso de 95% para 99%, o *upper bound* (limite superior) aumenta, "subindo", acontecendo o contrário com o *lower bound* (limite inferior) que diminui descendo, aumentando desta forma o próprio IC e consequentemente a região sombreada. Por outro lado, quando se diminui o nível de confiança, o *upper bound* diminui e o *lower bound* aumenta, diminuindo a região abrangida pelo intervalo de confiança.

Como referido anteriormente, o nível de confiança corresponde à probabilidade do valor observado se encontrar dentro do IC, isto é, para um nível de confiança de 95%, significa que existe 95% de probabilidade de o valor observado estar dentro do IC de previsão.

Como foi provado através dos modelos acima descritos, à medida que se aumenta o nível de confiança, a área do IC aumenta. Assim, pretendeu-se aferir se com este aumento do nível de confiança, a probabilidade do valor observado se encontrar dentro do IC também aumenta. Neste seguimento, geraram-se dois modelos, recorrendo ao ignorar último, com os seguintes parâmetros e *outputs* ilustrados na **Figura 25** e na **Figura 26**, respetivamente.

Modelo 5) Parâmetros:
HT= 12 meses
IL= 10 meses
IC= 75%
Saz.= 12 meses

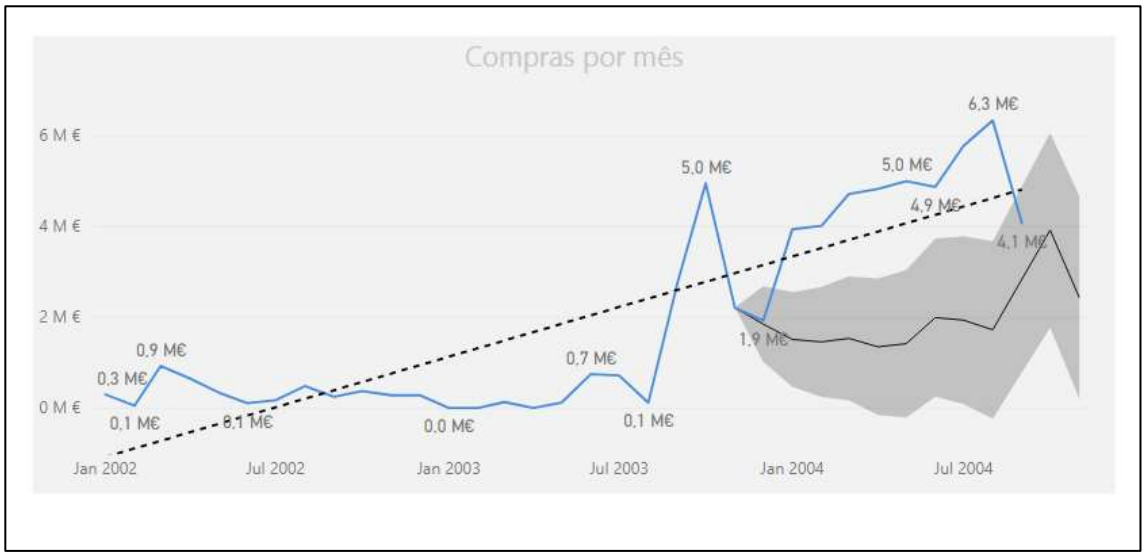

Figura 25 - Modelo 5 do teste ao parâmetro IC

Modelo 6) Parâmetros:
HT= 12 meses
IL= 10 meses
IC= 99%
Saz.= 12 meses

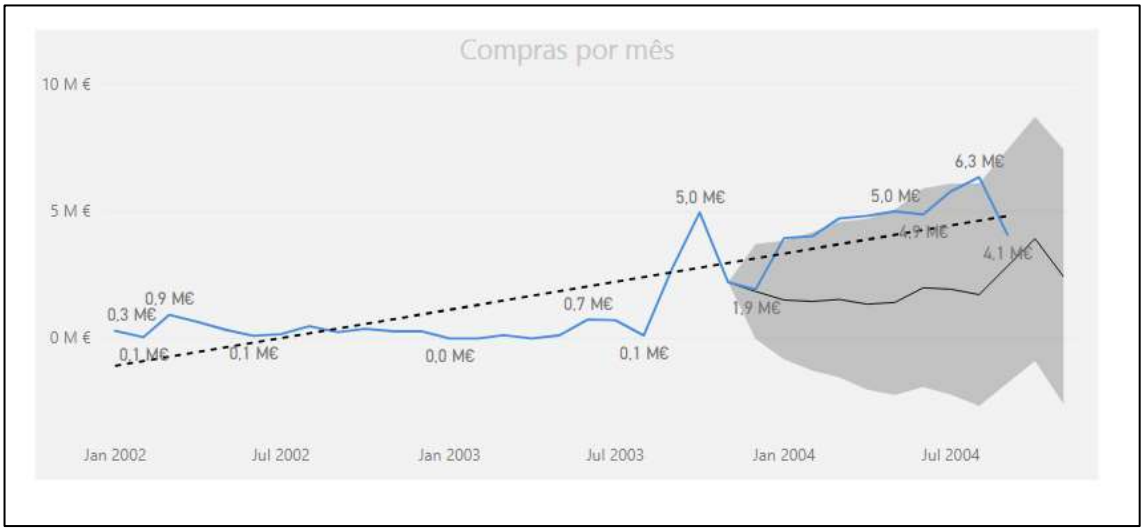

Figura 26 - Modelo 6 do teste ao parâmetro IC

No modelo 5, verifica-se que os valores observados se encontram fora da região sombreado, ao contrário do modelo 6, em que apesar de estarem no limite, os valores observados encontram-se quase sempre dentro da região em questão.

A partir destes, comprova-se que, com o aumento do nível de confiança, menor é margem de erro, o que significa que o *upper bound* aumenta e por isso "sobe", o *lower bound* diminui, "descendo", aumentando assim, a região abrangida pelo IC de previsão e, adicionalmente, a probabilidade do valor observado se encontrar dentro deste.

4.3.2.3 Teste ao parâmetro sazonalidade (saz.)

Procedeu-se a um teste ao parâmetro Sazonalidade (Saz.), com o objetivo de se entender o impacto e importância que a variação deste parâmetro tem na análise preditiva. Assumiu-se que os parâmeros restantes se mantêm constantes. Posto isto, confrontaram-se os diversos modelos gerados, onde o significado do HT, IL, IC e Saz, se mantêm iguais.

Modelo 1)	**Parâmetros:**
	HT= 24 meses
	IL= 0
	IC= 95%
	Saz.= 12 meses

Figura 27 - Modelo 1 do teste ao parâmetro sazonalidade

Modelo 2) Parâmetros:
HT= 24 meses
IL= 0
IC= 95%
Saz.= 6 meses

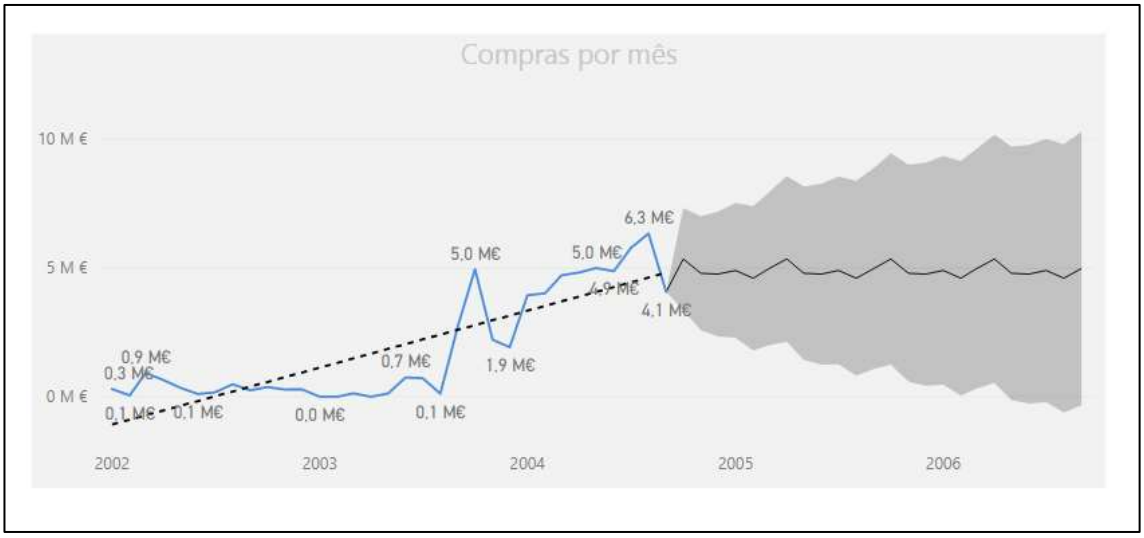

Figura 28 - Modelo 2 do teste ao parâmetro sazonalidade

Modelo 3) Parâmetros:
HT= 24 meses
IL= 0
IC= 95%
Saz.= 24 meses

Figura 29 - Modelo 3 do teste ao parâmetro sazonalidade

Modelo 4)	**Parâmetros:**
	HT= 24 meses
	IL= 0
	IC= 95%
	Saz.= 0 meses

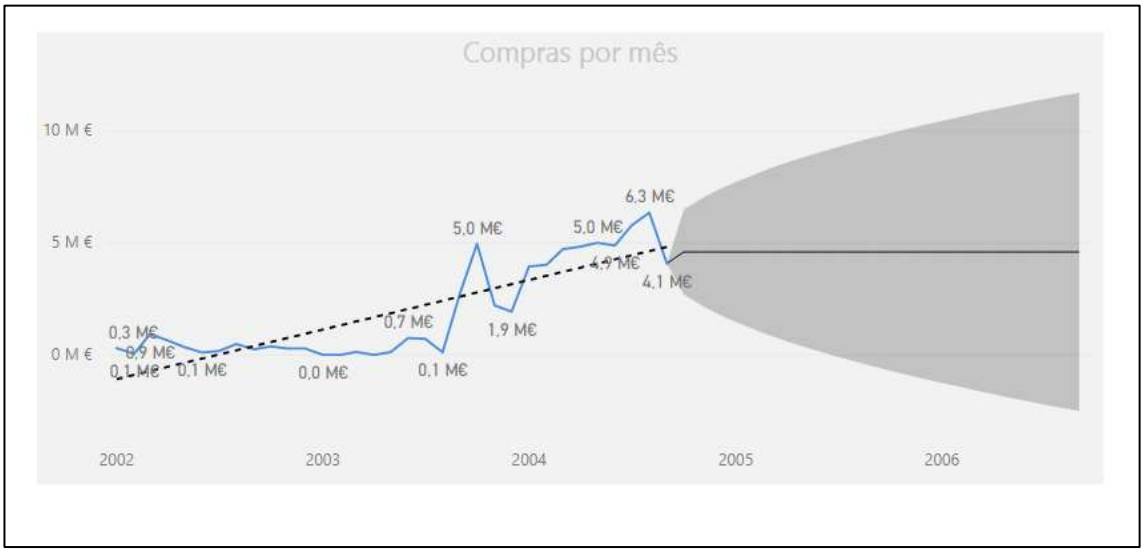

Figura 30 - Modelo 4 do teste ao parâmetro sazonalidade

Como se ilustra, na **Figura 27**, no modelo 1, para uma sazonalidade de 12 meses, o algoritmo utiliza a informação disponível dos últimos 12 meses e, para os 24 meses de previsão, projeta 2 períodos idênticos de 12 meses, reproduzindo assim o pico ocorrido no mês 10.

No modelo 2 ilustrado na **Figura 28**, e atribuindo uma sazonalidade de 6 meses, o algoritmo utiliza a informação disponível dos últimos 6 meses e, para os 24 meses de previsão, projeta 4 períodos idênticos de 6 meses, reproduzindo assim o pico ocorrido no mês 10 dividido em 2 momentos distintos, nomeadamente para o mês 10 e mês 4, perfazendo 4 picos no total.

No modelo 3 ilustrado na **Figura 29**, para uma sazonalidade de 24 meses, as projeções do algoritmo são muito semelhantes ao comportamento passado, ainda que os valores sejam ligeiramente diferentes.

O modelo 4, representado na **Figura 30**, para uma sazonalidade de 0 meses, verifica-se que, na ausência deste *input*, a previsão traduz-se numa linha reta, sem qualquer tipo de oscilação.

Este último modelo mostra a importância deste parâmetro no contexto de análise preditiva, uma vez que, é crucial que se tenha conhecimento relativo ao negócio e às suas especificidades para que, de forma manual, se consiga atribuir uma sazonalidade assertiva, alinhada com a realidade, de maneira a potenciar o desempenho do algoritmo.

A sazonalidade, é assim, um parâmetro de especial relevância, já que, como se pode constatar, o algoritmo de previsão utilizado, projeta e reproduz comportamentos futuros muito semelhantes aos passados, em função do período de sazonalidade definido.

Capítulo 5
Conclusão

5. Síntese da investigação, limitações e trabalho futuro

5.1 Síntese da investigação

O principal objetivo desta investigação foi a avaliação do potencial do *Power BI* para a análise preditiva, acessível a utilizadores genéricos. Isto é, utilizadores sem conhecimentos aprofundados de um leque de disciplinas, tais como matemática, estatística e informática ou, muitas vezes referidas como *Data Science*.

Em certa medida, passou por analisar o potencial que o *Power BI*, num contexto simplificado, traz à análise preditiva, tal como o *Excel* trouxe à estatística das organizações, com o fim de ajudar a perceber o que aconteceu, ou mais recentemente, o que está a acontecer na unidade económica com ou sem fins lucrativos. Em suma, aferir as potencialidades que a ferramenta PBI, na sua versão elementar, disponibiliza num contexto de análise preditiva.

O *Power BI* tem uma dupla potencialidade que consiste, por um lado na capacidade de visualização persuasiva, dinâmica, em tempo real e em múltiplas plataformas de dados. Por outro lado, as suas capacidades analíticas e preditivas, com a versatilidade de múltiplos conectores a múltiplas fontes de dados. Tudo isto se destina a diferentes níveis de utilizadores e a diferenciados cenários tecnológicos dos mais simplificados aos mais complexos.

Partindo da análise preditiva efetuada, o PBI é uma ferramenta que reúne várias funcionalidades interessantes e que, para o propósito e contexto respondeu muito bem. É um *software* de fácil utilização, que disponibiliza meios

para se efetuar uma análise preditiva avançada, de forma simples, rigorosa e intuitiva. Concede, ainda, aos seus utilizadores a aplicação de inteligência artificial de forma simples aos seus dados, dando suporte aos processos de tomada de decisão, sem que os mesmos tenham de ser conhecedores profundos da BD. Porém, é também uma ferramenta dotada de linguagens complexas que permitem a qualquer utilizador obter um maior nível de conhecimento aplicado a contextos de análises avançadas.

Na prática, e após terem sido testados cenários simplificados aos diversos parâmetros, conclui-se que estes são determinantes para o bom desempenho do algoritmo que a ferramenta utiliza. Verifica-se, assim, que o desempenho do algoritmo, bem como a qualidade de previsão inerente, estão dependentes de uma correta configuração destes parâmetros, sendo que uma configuração incorreta poderá comprometer a qualidade de previsão e, por conseguinte, os resultados da mesma.

Contudo, existem determinadas limitações associadas, sendo que neste caso os dados que serviram de base ao presente estudo, aparentam ser provenientes de uma empresa não madura, em fase de implementação, pelo facto de não haver registo de compras em muitos dos meses iniciais e uma elevada discrepância entre estes e os meses em que existem. Assim, esta inconsistência dos dados, compromete a qualidade de previsão, no sentido em que, são bastante irregulares e com um histórico curto. Adicionalmente, o facto de o algoritmo utilizado pelo *Power BI Desktop* ser ainda desconhecido, constitui outra limitação, já que se assumiu à *priori* que este implementaria o mesmo utilizado pelo *Power View*.

Ao longo deste estudo, conseguiu-se dar resposta à questão de investigação que aqui recordo: "Qual o potencial do *Power BI Desktop* para a Análise Preditiva?" assim como aos vários objetivos formulados para avaliar o seu potencial aplicado que se materializam no **Quadro 4** abaixo.

Objetivos	Conclusão
Identificar recursos e potencialidades disponíveis no *Power BI* para o efeito;	Para o propósito, identificou-se o painel de análise, bem como o recurso ao R e do *Azure Machine Learning;*
Pesquisar recursos adicionais integráveis no *Power BI* para análise preditiva;	*O MarketPlace* serve o objetivo, já que, recorrendo a esta funcionalidade, é possível efetuar análises avançadas e descarregar visualizações incomuns, criativas e únicas;
Identificar competências essenciais para aplicar em contexto de *Power BI* na análise preditiva;	Identificaram-se os requisitos determinantes à realização da análise preditiva, nomeadamente ao que ao formato dos dados diz respeito;
Explorar experimentalmente os recursos, de forma a compreender, na prática, o impacto e contributo dos mesmos neste tipo de análise;	Constituindo o *core* deste estudo, exploram-se vários cenários recorrendo para isso, à testagem dos parâmetros de configuração considerados fulcrais;
Avaliar o rigor, qualidade e suporte que a ferramenta poderá conferir aos seus utilizadores no âmbito da previsão;	No seguimento da exploração dos parâmetros acima mencionados, permitiram avaliar o rigor, qualidade e suporte que o algoritmo utilizado confere neste tipo de análise;

Detetar benefícios, fragilidades e desafios no âmbito da previsão e do *Power BI*.	As vantagens inerentes à adesão desta ferramenta de BI são mencionadas ao longo do estudo. As limitações associadas aos dados, bem como a incerteza inerente ao algoritmo utilizado constituem fragilidades. Por último, o trabalho futuro ilustra um dos desafios lançados.

Quadro 4 – Conclusão por objetivo proposto

Para recordar, o site pode ser consultado aqui, que ilustra de forma multimédia, o caminho a seguir para a realização de uma análise preditiva no contexto e destinatários descritos.

5.2 Trabalho futuro

Sendo este um estudo de exploração num contexto académico, não foi possível conduzi-lo e implementá-lo num contexto *corporate* nas mais diversificadas organizações e setores, não conseguindo, fazer assim, uso da versão *Premium* da ferramenta disponível pela MS. Acreditando no seu potencial e benefícios inerentes, sugere-se desta forma que, futuramente, se equacione experimentar e tirar ilações do potencial desta ferramenta num formato mais formal e orientado para profissionais na área.

Referências Bibliográficas

Acito, F., & Khatri, V. (2014). Business Analytics: Why now and what next? Business Horizons.

Alves, C. (2019). *Quadrante Mágico 2019: Power BI segue líder pela 12º vez segundo o Gartner*. Obtido em 07 de 04 de 2020, de bi9: https://blog.bi9.com.br/quadrante-magico-2019-power-bi/

Antonelli, R. A. (2009). Conhecendo o Business Intelligence (BI) - Uma Ferramenta de Auxílio à Tomada de Decisão.

AutoML. (2020). *O que é aprendizagem automática de máquinas (AutoML)?* Obtido em 25 de 04 de 2020, de Microsoft: https://docs.microsoft.com/pt-pt/azure/machine-learning/concept-automated-ml

Azevedo, A., & M.F.Santos. (2008). KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW.

Barbieri, C. (2001). *BI - Business Intelligence: modelagem e tecnologia.* Axacel Books.

Berry, M. J., & S.Linoff, G. (2004). *Data Mining Techniques.*

Bhattacharjee, J. (2019). *Common metrics for Time Series Analysis*. Obtido em 27 de 04 de 2020, de Medium: https://medium.com/@joydeepubuntu/common-metrics-for-time-series-analysis-f3ca4b29fe42

BI, P. (2019). *O que é o Power BI*. Obtido em 25 de 02 de 2020, de Microsoft : https://docs.microsoft.com/pt-pt/power-bi/fundamentals/power-bi-overview

Blast. (2019). *Microsoft Power BI Consulting*. Obtido em 10 de 05 de 2020, de Blast: https://www.blastanalytics.com/power-bi-consulting

Borges, M., Cardozo, C., & Filho, O. (2018). DOS DADOS AO CONHECIMENTO: BUSINESS INTELLIGENCE COMO FERRAMENTA PARA APOIO À TOMADA DE DECISÃO.

Bouwens, L. (2018). *The five user types of Power BI reporting & analytics*. Obtido em 25 de 04 de 2020, de MSBIBlog: https://www.msbiblog.com/2018/08/28/the-five-user-types-of-power-bi-reporting-analytics/

Brownlee, J. (2020). *A gentle Introduction to Exponential Smoothing for Time Series Forecasting in Phyton*. Obtido em 29 de 04 de 2020, de Machine Learning Mastery: https://machinelearningmastery.com/exponential-smoothing-for-time-series-forecasting-in-python/

Burki, I. (2018). *Take Advantage of Forecasting in Power BI with the Analytics Pane*. Obtido em 12 de 05 de 2020, de Stoneridge Software : https://stoneridgesoftware.com/take-advantage-of-forecasting-in-power-bi-with-the-analytics-pane/

Camilo, C. O., & Silva, J. C. (2009). *Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas.*

Castro, L. M., & Silva, M. A. (s.d.). Business Intelligence (BI): Análise comparativa entre as ferramentas líderes no mercado.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0-Step-by-step data mining guide.

Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An Overview of Business Intelligence Technology.

Daderman, A., & Rosander, S. (2018). Evaluating Frameworks for Implementing Machine Learning in Signal Processing: A Comparative Study of CRISP-DM, SEMMA and KDD.

Daitan. (2019). *Exponential Smoothing Methods for Time Series Forecasting*. Obtido em 28 de 04 de 2020, de Medium: https://medium.com/better-programming/exponential-smoothing-methods-for-time-series-forecasting-d571005cdf80

Delen, D. (2015). *Real-World Data Mining: Applied Business Analytics and Decision Making.* Pearson Education, Inc.

E.Adam, E., & Ebert, R. J. (1991). Unit 4: Need and Importance Forecast. Em *Production and Operations Management: Concepts, Models, and Behavior* (pp. 5-11).

Elena, C. (2011). Business Intelligence. *Journal of Knowledge Management, Economics and Information Technology*.

Evans, J. R. (2016). *Business Analytics : Methods, Models and Decisions.* Pearson.

Fávero, L. P. (2019). *KDD e Data Mining: mais do que apenas dois conceitos*. Obtido em 30 de 01 de 2020, de https://www.itforum365.com.br/colunas/kdd-e-data-mining-mais-do-que-apenas-conceitos/

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data.

Ferrari, A., & Russo, M. (2016). *Introducing Microsoft Power BI.* Microsoft Corporation .

FluentPro. (2019). *Portfolio Forecasting with Microsoft Power BI: How-To Guide*. Obtido em 28 de 04 de 2020, de FluentPro: Portfolio Forecasting with Microsoft Power BI: How-To Guide

Gama, J., Carvalho, A. d., Faceli, K., Lorena, A. C., & Oliveira, M. (2015). *Extração de Conhecimento de Dados - Data Mining.* Edições Sílabo.

Gandomi, A., & Haider, M. (2014). Beyond the hype: Big data concepts, methods, and analyticsAmir. p. 7.

Gartner. (s.d.). *Gartner Magic Quadrant*. Obtido em 07 de 04 de 2020, de Gartner: https://www.gartner.com/en/research/methodologies/magic-quadrants-research

Gestão, P. (2017). *Power BI e Business Intelligence: os melhores aliados para as suas decisões*. Obtido em 27 de 02 de 2020, de Portal de Gestão: https://www.portal-gestao.com/artigos/7931-power-bi-e-business-intelligence-os-melhores-aliados-para-as-suas-decis%C3%B5es.html

Goldschmidt, R., & Passos, E. (2005). *Data Mining: um guia prático* . Campus.

Guilfoyle, P. (2017). Forecasting in Power BI. Obtido em 12 de 05 de 2020, de https://www.youtube.com/watch?v=XIlPkyyztho

Han, J., Pei, J., & Kamber, M. (2012). *Concepts and Techniques, 3rd ed.* ELSEVIER.

Hand, D. J. (2007). Principles of Data Mining.

Hevner, A., T.March, S., Park, J., & Ram, S. (2004). Design Science in Information Systems Research.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Paractice*. Obtido em 30 de 04 de 2020, de https://otexts.com/fpp2/data-methods.html

Institute, S. (2006). *Enterprise Miner - SEMMA*. Obtido em 02 de 02 de 2020, de SAS: http://faculty.smu.edu/tfomby/eco5385_eco6380/data/SPSS/SAS%20_%20SEMMA.pdf

Iseminger, D. (2019). *O que é o Power BI Desktop?* Obtido em 15 de 02 de 2020, de Microsoft: https://docs.microsoft.com/pt-pt/power-bi/fundamentals/desktop-what-is-desktop

Iseminger, D. (2020). *Introdução ao Power BI Desktop*. Obtido em 30 de 01 de 2020, de Microsoft: https://docs.microsoft.com/pt-pt/power-bi/fundamentals/desktop-getting-started

Kumar, V., & M.L.Garg. (2018). Predictive Analytics: A Review of Trends and Techniques.

Laruccia, M. M., Silva, R. S., & Chiarelli, G. D. (2013). Discussão sobre o Business Intelligence em empresas de Tecnologia da Informação.

Lima, J. R., & Capitão, Z. (2003). e-Learning e e-Conteúdos . INOVA.

Louzada, P. (2019). *BLOG ANÁLISE DE DADOS*. Obtido em 26 de 05 de 2020, de Quais são as 7 versões do Power BI? Como elas funcionam?: https://www.fm2s.com.br/quais-sao-as-7-versoes-do-power-bi-como-elas-funcionam/

Ludovice, C. (2017). *O que é e quais são as vantagens de usar o Power BI?* Obtido em 01 de 03 de 2020, de IMPACTA: https://www.impacta.com.br/blog/2017/06/12/o-que-e-e-quais-sao-os-beneficios-de-usar-o-power-bi/

Luhn, H. (1958). A Business Intelligence System. *IBM Journal of Research and Development*.

Malheiro, S. (2020). *Power BI - Uma dádiva para os Analistas!* Obtido em 19 de 02 de 2020, de Portal de Gestão: https://www.portal-gestao.com/artigos/8112-power-bi-uma-d%C3%A1diva-para-os-analistas.html

Martinez, M. (2016). *The Alteryx Starter Kit for Microsoft.* Obtido em 11 de 03 de 2020, de Microsoft: https://powerbi.microsoft.com/en-us/blog/the-alteryx-starter-kit-for-microsoft/

Mehta, S. (2017). *Data Forecasting and Analytics with Power BI Desktop*. Obtido em 12 de 05 de 2020, de MSSQL TIPS: https://www.mssqltips.com/sqlservertip/5085/data-forecasting-and-analytics-with-power-bi-desktop/

Microsoft. (2014). *Describing the forecasting models in Power View*. Obtido em 27 de 04 de 2020, de Microsoft Power BI Blog:

https://powerbi.microsoft.com/en-us/blog/describing-the-forecasting-models-in-power-view/

Microsoft. (2014). *Introducing new forecasting capabilities in Power View for Office 365*. Obtido em 27 de 04 de 2020, de Microsoft Power BI Blog: https://powerbi.microsoft.com/en-us/blog/introducing-new-forecasting-capabilities-in-power-view-for-office-365/

Microsoft. (2018). *O que é o Power BI Embedded no Azure?* Obtido em 03 de 04 de 2020, de Microsoft: https://docs.microsoft.com/pt-pt/power-bi/developer/embedded/azure-pbie-what-is-power-bi-embedded

Microsoft. (2019). *Aplicar as noções básicas do DAX no Power BI Desktop*. Obtido em 28 de 04 de 2020, de Microsoft: https://docs.microsoft.com/pt-pt/power-bi/transform-model/desktop-quickstart-learn-dax-basics

Microsoft. (2019). *O que é Power BI?* Obtido em 05 de 05 de 2020, de Microsoft: https://docs.microsoft.com/pt-pt/power-bi/fundamentals/power-bi-overview

Microsoft. (2019). *What is an on-premises data gateway? (O que é um gateway de dados no local?* Obtido em 15 de 04 de 2020, de Microsoft: https://docs.microsoft.com/pt-pt/power-bi/connect-data/service-gateway-onprem

Microsoft. (2020). *O que é o Power BI Report Server?* Obtido em 16 de 04 de 2020, de Microsoft: https://docs.microsoft.com/pt-pt/power-bi/report-server/get-started

Microsoft. (s.d.). *Advanced Analytics with Power BI.* Obtido em 09 de 04 de 2020, de Microsoft: https://www.arbelatech.com/insights/white-papers/advanced-analytics-with-power-bi

Microsoft. (s.d.). *Comparação de funcionalidades do Power BI*. Obtido em 15 de 05 de 2020, de Microsoft: https://powerbi.microsoft.com/pt-pt/pricing/

Moigne, J. L. (1978). La théorie du système d'information organisationnel .

Nabler. (s.d.). *The impact of Power BI in Business Intelligence*. Obtido em 27 de 02 de 2020, de Nabler: https://www.nabler.com/articles/the-impact-of-power-bi-in-business-intelligence/

Negash, S., & Gray, P. (2003). Business Intellogence.

Oliveira, A. (2009). *Informação e Sistemas de Informação.* Refer Telecom.

Olszak, C. M., & Ziemba, E. (2007). Approach to Building and Implementing Business Intelligence Systems.

Palmer, T. (2016). *Visualize your Prevedere data with Power BI.* Obtido em 11 de 03 de 2020, de Microsoft: https://powerbi.microsoft.com/en-us/blog/visualize-your-prevedere-data-with-power-bi/

Parashar, S. (2019). *Predictive Analytics using Power BI*. Obtido em 10 de 04 de 2020, de Microsoft : https://community.dynamics.com/365/b/businesstransformationdynamics365/posts/predictive-analytics-using-power-bi

Pereira, M. (2020). *Power BI: O que é e para que serve*. Obtido em 26 de 02 de 2020, de Voitto: https://www.voitto.com.br/blog/artigo/o-que-e-power-bi

Powell, B. (2017). *Microsoft Power BI Cookbook: Creating Business Intelligence Solutions of Analytical Data Models, Reports, and Dashboards.* Packt.

Powell, B. (2018). *Mastering Microsoft Power BI:Expert techniques for effective data analytics and business intelligence.* Packt.

Primak, F. V. (2008). *Decisões com B.I. - Business Intelligence.* Ciencia Moderna.

Rad, R. (2019). *From Rookie to Rock Star* (Vol. I). RADACAD Systems Limited.

Ranjan, J. (2005). Business Intelligence: concepts, components, techniques and benefits. *Journal of Theoretical and Applied Information Technology*.

Reigeluth, C. M. (2009). Intructional-Design Theories and Models : A New Paradigm of Instructional Theory. Routledge.

Richardson, J., Sallam, R., Schlegel, K., Kronz, A., & Sun, J. (2020). *Microsoft*. Obtido em 09 de 05 de 2020, de 2020 Gartner Magic Quadrant for Analytics and Business Intelligence Platforms: https://info.microsoft.com/ww-landing-2020-gartner-magic-quadrant-for-analytics-and-business-intelligence.html?LCID=EN-US&ls=Website

Santos, M., & Ramos, I. (2017). *Business Intelligence - Da informação ao conhecimento.* FCA.

Scardina, J., & Horwitz, L. (s.d.). *Microsoft Power BI*. Obtido em 03 de 06 de 2020, de Search Content Management: https://searchcontentmanagement.techtarget.com/definition/Microsoft-Power-BI

Sezões, C., Oliveira, J., & Baptista, M. (2006). *Business Intelligence.* SPI - Sociedade Portuguesa de Inovação.

Sikes, N. (2017). *An Introduction to Self-Service Business Intelligence*. Obtido em 12 de 02 de 2020, de DOMO: https://www.domo.com/blog/introduction-to-self-service-business-intelligence/

Silva, M. d. (2017). Business Analytics na Prática. *Vida Económica*.

Singh, K., Shastri, S., Bhadwal, A. S., Kour, P., Kumari, M., Sharma, D., & Mansotra, P. (2019). Implementation of Exponential Smoothing for Forecasting Time Series Data.

Smoak, A. B. (2019). *How to Generate a Forecast in Power BI*. Obtido em 28 de 04 de 2020, de https://anthonysmoak.com/2019/02/02/how-to-generate-a-forecast-in-power-bi/

Soares, A. C., & Silva, P. D. (s.d). Análise de Ferramentas de Business Intelligence com destaque dos serviços de BI na Cloud Computing.

Sonna, D. (2018). *Ferramentas de BI: Conheça as ferramentas de BI mais utilizadas no mercado*. Obtido em 26 de 02 de 2020, de SONNA: https://www.sonna.com.br/ferramentas-de-bi-conheca-as-ferramentas-de-bi-mais-utilizadas-no-mercado/

T.Moss, L., & Atre, S. (2003). *The Complete Project Lifecycle for Decision-Support Applications.* Pearson Education, Inc.

Tkachuk, R. (2019). *Power BI Premium and Azure Analysis Services*. Obtido em 10 de 02 de 2020, de Blogue do Microsoft Power BI: https://powerbi.microsoft.com/pt-pt/blog/power-bi-premium-and-azure-analysis-services/

Ulag, A. (2019). *Announcing New AI and Enterprise features for Power BI.* Obtido em 11 de 03 de 2020, de Microsoft Power BI Blog: https://powerbi.microsoft.com/en-us/blog/announcing-new-ai-and-enterprise-features-for-power-bi/

Valentim, M. L. (2002). Inteligência Competitiva em organizações: dado, informação e conhecimento.

Vasconcellos, P. (2017). *CRISP-DM, SEMMA e KDD: conheça as melhores técnicas de exploração de dados*. Obtido em 29 de 01 de 2020, de Medium: https://paulovasconcellos.com.br/crisp-dm-semma-e-kdd-conhe%C3%A7a-as-melhores-t%C3%A9cnicas-para-explora%C3%A7%C3%A3o-de-dados-560d294547d2

Velosio. (2019). *ERP Software Blog.* Obtido em 16 de 03 de 2020, de PowerBI License and User Types: What Skills are Needed for Developers,

Analysts, and End Users: https://www.erpsoftwareblog.com/2019/09/powerbi-license-types-developers-analysts-end-users/

Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making.* John Wiley & Sons Ltd.

Wade, C. (2019). *Large models in Power BI Premium public preview*. Obtido em 10 de 02 de 2020, de Microsoft Power BI Blog: https://powerbi.microsoft.com/en-us/blog/large-models-in-power-bi-premium-public-preview/

Wannapipat, W., Chaijaroen, S., & Somabut, A. (2013). SOI Model of Learning to Tablet Instructional Implication Generating .

Zorrinho, C. (1991). *Gestão da informação.* Editorial Presença.

Apêndice A
Business Intelligence

A.1 Evolução do conceito de BI

Turban e Volonimo encontram-se alinhados com os restantes autores, definindo o *BI* como um conjunto de ferramentas, metodologias e aplicações que funcionam como suporte ao processo de decisão de uma organização, destacando-lhe três funções principais: consultar, analisar e relatar (Turban & Volonimo, 2013, p. 330).

Em 2009, surge **Vercellis** ao afirmar que o *BI* pode ser definido como um "apoio de modelos matemáticos e metodologias de análise que explorem os dados disponíveis para gerar informação e conhecimento para processos de tomada de decisões complexas" (Vercellis, 2009).

Segundo a **TDWI** (*Tranforming Data With Intelligence*), uma empresa fundada em 1995, vê o *BI* como um processo de associação de dados, tecnologia e conhecimento humano com o objetivo principal de melhorar a tomada de decisão e conseguir orientar as empresas para o sucesso, ou seja, o *BI* é uma combinação de dados, tecnologia, análise e conhecimento humano com o propósito de otimizar as decisões negociais (TDWI, 2013).

A **Forrester** define o BI como "um conjunto de metodologias, processos, arquiteturas e tecnologias que influenciam o retorno dos processos de gestão de informação para análises, relatórios, gestão de *performance* e entrega de informação (Forrester, 2015).

Para **Mircea e Andreescu**, as ferramentas de *BI* funcionam como auxílio e suporte na compreensão de fatores que influenciam indicadores e métricas de

desempenho, acabando por, em paralelo, ajudar os gestores a escolherem as informações de forma assertiva para a prática de uma boa gestão (Mircea e Andreescu, 2011).

Larissa Moss define o *BI* como sendo uma arquitetura em oposição a um produto ou sistema, sendo esta constituída por um conjunto de aplicações integradas tanto operacionais como de suporte à decisão, englobando todas as BD necessárias que potenciem o acesso simples e rápido a toda a informação existente (T.Moss & Atre, 2003).

Para Angeloni e **Reis,** o conceito de *BI* refere-se ao conjunto de "metodologias de gestão implementadas através de ferramentas de *software*, cuja função é proporcionar ganhos nos processos decisórios gerenciais e da alta administração nas organizações, baseada na capacidade analítica das ferramentas que integram num só lugar todas as informações necessárias ao processo decisório". Reforça ainda a ideia de que o core do *BI* é o de transformar dados em informações de qualidade e estas em conhecimento que sustentabilizem o processo de tomada de decisão, com o objetivo de gerar vantagem competitiva (Antonelli, 2009).

A.2 ETL - Definições, objetivos e características

Antes de se iniciar ou desenvolver um processo de ETL, deverá efetuar-se um levantamento relativo a aspetos fundamentais, uma vez que estes fatores condicionarão a otimização de todo o processo de ETL, evitando desta forma problemas em fases já finais de projetos, como casos de inconsistências nos dados, *missing values*, redundância de dados, entre outros entraves que podem comprometer todo o processo (Sezões *et al*., 2006). Assim, se for efetuada uma análise atempada, evitam-se estes constrangimentos e despesas sem fundamento. Contudo, já existem processos de ETL que incorporam este tipo de análise (Sezões *et al*., 2006).

De uma perspetiva mais específica, o processo de ETL refere-se a todo o conjunto de ferramentas e processos de Extração, Transformação e Carregamento de dados cujo principal objetivo é a homogeneização dos dados, a sua limpeza e o respetivo carregamento para o *Data Warehouse* (Vassiliadis, Simitsis et al, 2002), (Sezões *et al*., 2006). Para que tal aconteça, é necessário que as organizações disponham de uma área de trabalho apropriada, bem como de estruturas de dados e de todo um conjunto de processos que permitam o acesso a dados-fonte e às suas funções principais (Santos & Ramos, 2017).

A arquitetura e planeamento de um processo de ETL deverão ser versáteis, no sentido em que deve ser extensível a diversas áreas, já que este se caracteriza por ser eminentemente sequencial (isto é, não se podem carregar as vendas sem carregar os produtos e os clientes – as dimensões vêm sempre antes dos factos), portanto, é obrigatório que se tenham em conta estes fatores na sua arquitetura (Sezões *et al*., 2006).

Relativamente à sua arquitetura, esta divide-se em cinco etapas principais: extração de dados, limpeza dos dados, transformação de dados, carregamento e por último, a fase de atualização dos dados (Barbieri, 2001), (Santos & Ramos, 2017).

A extração dos dados, tal como o nome indica, refere-se à recolha dos dados de diversas fontes através dos procedimentos adequados, permitindo a sua organização em diversos conjuntos de dados. Os dados extraídos são temporariamente transferidos e armazenados na área de estágio denominada por *Data Staging Area* (DSA), onde se vão aplicar os processos de limpeza e transformação de dados respetivamente. É nestes dois processos que, por um lado, se identificam possíveis erros ou imperfeições nos dados e se efetua a sua correção sempre que possível para, posteriormente, se padronizarem e homogeneizarem os dados para um formato ideal, único e isentos de erros, de forma a serem devidamente depositados no *DW*. Seguidamente, os dados existentes na *DSA* são inseridos no *DW*, dando-se a etapa de carregamento. O

carregamento engloba tarefas como ordenação, agregação, consolidação, verificação da integridade dos dados, entre outras. Finalmente, a última etapa é a atualização dos dados que ocorre sempre que existam novos dados nas fontes ou quando os dados já existentes sofrem qualquer tipo de alteração e, como tal, tenham de ser carregados para o *DW* (Barbieri, 2001), (Santos & Ramos, 2017), (Vercellis, 2009).

O processo de ETL é considerado crucial e uma das etapas mais críticas na construção e desenvolvimento de um *Data Warehouse*, uma vez que é a partir dos dados previamente tratados e transformados que é possível efetuar a sua posterior análise e tirar ilações (Antonelli, 2009), (Primak, 2008).

Tendo em conta a sua importância e nível de complexidade, a concretização deste processo tende a consumir uma quantidade avultada de tempo, chegando a atingir 70% dos recursos necessários para a implementação e manutenção de um *Data Warehouse*, onde, se atribui primazia à etapa de transformação dos dados (Sezões *et al*., 2006), (Santos & Ramos, 2017).

Contudo, é de salientar que os benefícios gerados por um processo de ETL serão significativamente visíveis a longo prazo, estando o aumento da sua produtividade diretamente relacionado e dependente da qualidade que os dados apresentam (Primak, 2008), (Kimball e Caserta, 2004), (Santos & Ramos, 2017).

A.3 Data Warehouse e Data Mart

Como forma introdutória, é possível afirmar que uma base de dados (BD) é um conjunto de dados organizados de uma forma sistematizada. São vários os tipos/modelos de bases de dados, realçando-se o tipo relacional e o dimensional, sendo este último o que maioritariamente se utiliza num *DW* (Sezões *et al*., 2006).

Um modelo dimensional assemelha-se a um cubo constituído por três ou mais dimensões, onde cada uma representa um atributo distinto, potenciando desta forma, uma melhor visualização dos dados devido à sua organização

simplificada dos dados. Além de que também contribui para flexibilizar o processamento de consultas, e dispõe de versatilidade para que eventuais alterações possam ser feitas ao modelo (Kimball et al, 1998), (Ross, 2002), (Sezões *et al*., 2006).

A par das bases de dados relacionais, estas ocorrem quando se dá uma junção de tabelas através de um campo comum a ambas. Exemplificando: se existirem duas tabelas – clientes e moradas, com campos em comum, e se efetuar a sua junção, vai resultar numa BD sem informação redundante. Este processo designa-se normalização, que é bastante habitual, tendo em conta que quanto mais normalizada uma BD estiver, mais eficiente serão efetuadas as atualizações das tabelas. A exceção aplica-se em consultas complexas de várias tabelas, como é o caso dos sistemas OLAP, uma vez que a rapidez da sua consulta aumenta quanto mais desnormalizada se apresentar a BD (Sezões *et al*., 2006).

Por definição, um *Data Warehouse* consiste num armazém de dados, semelhante a um repositório integrado de dados especificamente construído para o armazenamento e consolidação de toda a informação considerada relevante pelas organizações, num formato-padrão completamente desprovido de erros, válido e consistente, que visa facilitar a sua análise e posterior tomada de decisão (Santos & Ramos, 2017). Toda a informação que é armazenada num *Data Warehouse* é devidamente rotulada de cariz temporal, cuja capacidade de armazenamento excede múltiplos anos (Santos & Ramos, 2017). Um *Data Warehouse* é uma base de dados dimensional que é mantida e alimentada de forma autónoma face às bases de dados operacionais (Santos & Ramos, 2017).

Citando Inmon (1996b) relativamente ao conceito de DW – "Um DW consiste num conjunto de dados orientado por assunto, integrado, catalogado temporalmente e não volátil que suporta os gestores no processo de tomada de decisão" (Santos & Ramos, 2017).

A par da citação acima transcrita, é possível destacar e sistematizar cinco grandes características inerentes a qualquer *DW*:

- **Integrado** – Um *DW* deverá representar uma fonte única e expansionista construído a partir do cruzamento de informação proveniente de múltiplas fontes heterogéneas de dados. A integração de dados refere-se ao processo através do qual os dados-fonte são transformados e modificados, de forma a assegurar a sua consistência, que permita que sejam inseridos sob a forma desejada no mesmo (Santos & Ramos, 2017), (Sezões *et al*., 2006).
- **Orientados por assunto** – Num *DW*, os dados e a informação são organizados e divididos em compartimentos, de acordo com os principais segmentos de uma organização, como por exemplo, segundo clientes, fornecedores, produtos, entre outros. Esta divisão tem como objetivo agilizar e satisfazer as necessidades dos seus utilizadores, uma vez que estes sistemas facultam uma perspetiva simplificada relativamente a um dado assunto, eliminando os dados que sejam considerados não relevantes, fazendo desta forma uma espécie de triagem (Santos & Ramos, 2017), (Sezões *et al*., 2006).
- **Catalogado temporalmente** – O *core* dos *DW* reside em conseguir fornecer informações não só atuais, mas também, apresentar uma perspetiva histórica, variando assim no tempo. É, desta forma, capaz de produzir análises de evoluções históricas, cujo alcance temporal se encontra entre 5 e 10 anos (Santos & Ramos, 2017), (Sezões *et al*., 2006).
- **Não volátil** – Um *DW* armazena dados e informação bastante estável não podendo ser alterados ou eliminados após o seu carregamento. Assim, um *DW* é tipicamente responsável por duas tarefas: o carregamento dos dados de forma regular e o acesso aos mesmos para processamento de consultas. Desta forma, são sempre disponibilizados aos utilizadores dados corretos com elevada credibilidade (Santos & Ramos, 2017), (Sezões *et al*., 2006).
- **Acessíveis** – um dos propósitos essenciais da criação de um *DW* é conceder e agilizar o acesso a toda a informação relevante de uma forma

> simples e astuta. Como tal, o *DW* tem como principal fim o de conseguir, em tempo útil, responder às necessidades do negócio no que concerne à obtenção e análise de informação, transformando dados de diversas fontes em informações relevantes para a organização (Sezões *et al*., 2006).

A partir de toda esta informação, é possível então concluir que um DW é considerado um repositório valioso de dados históricos de natureza operacional, transacional e consistentes de uma organização agrupados numa BD que atua como um facilitador e suporte à tomada de decisões estratégicas para o negócio, como também prima pela sua análise e *reporting* (Santos & Ramos, 2017), (Sezões *et al*, 2006), (Han e Kamber, 2001).

São dois os motivos fulcrais que levam à criação e desenvolvimento de um DW por parte de uma organização. O primeiro refere-se à necessidade que as organizações têm de conseguir obter todos os dados e informações distribuídos em diferentes fontes integrados numa só BD, o que potencia desde logo uma melhor visão e análise global acerca do panorama organizacional. O outro fator prende-se com a necessidade que as organizações demonstram sobre efetuarem uma triagem em relação aos dados que são utilizados nas operações correntes dos que são utilizados para análise e *reporting* (Sezões *et al*., 2006).

Data Mart

Por definição e citando Sezões *et al*., um *Data Mart* pode ser considerado como "uma versão mais especializada e específica de um *Data Warehouse"*, isto é, acaba por ser também um repositório de dados homogéneos que estão de certa forma agrupados e alinhados por categorias.

Tanto um *DW* como um *Data Mart* partilham a mesma tecnologia, porém, enquanto que um *Data Warehouse* engloba todos os dados de uma organização, um *Data Mart* inclui apenas a informação relativa a um departamento integrante

da organização, como o departamento de Marketing, Recursos Humanos, entre outros (Sezões *et al*., 2006), (Vercellis, 2009).

Tal significa, que, um *Data Mart* pode ser visto como um *DW* departamental, de tamanho reduzido que fornece informações que servem de suporte a tomada de decisões, face a determinado departamento, em oposição a um *DW*, que por sua vez, disponibiliza informações que suportam decisões para a organização como um todo (Antonelli, 2009), (Vercellis, 2009).

Em paralelo, são vários os fatores que fundamentam a criação de *data marts* como é o caso de melhoria de desempenho, tendo em conta que ao efetuar-se a separação de um conjunto de dados provenientes de um *DW*, resultará numa maior eficiência no tratamento e processamento dos mesmos. Inequivocamente, este tipo de repositório mais reduzido leva a uma maior segurança face à informação da organização, já que, é possível fazer a distinção entre a informação autorizada e a informação confidencial e por último, também apresenta vantagens no que concerne à sua utilidade, na medida em que frequentemente as organizações sentem necessidade de obter um modelo de dados diferente do *DW* que possa ser aplicado a finalidades de negócio também diferentes (Sezões *et al*., 2006). Contudo, é importante ressalvar que a criação de *data marts* deverá ser feita com bastante consciência e corretamente justificada, porque caso contrário, poderá resultar em informação redundante, inconsistente e incoerente (Sezões *et al*., 2006).

Posto isto, dependendo da natureza organizacional, esta pode optar pela implementação de um *Data Warehouse* organizacional, *Data Marts* independentes ou de *Data Marts* dependentes do *Data Warehouse*, sendo que é a própria organização a fazer a escolha do modelo que melhor se adequa, após efetuar um levantamento das necessidades e avaliação das características (Santos & Ramos, 2017).

Se a escolha passar por um DW organizacional, este englobará e integrará os dados de toda a organização como já foi acima referido, sendo alimentado por

um ou mais sistemas operacionais, podendo estes serem complementados por dados externos (Santos & Ramos, 2017). Se, por outro lado, a organização optar por um modelo de *Data Marts* independentes, estes integram subconjuntos de dados da organização relevantes para determinados grupos de utilizadores como já foi referido anteriormente, sendo estes alimentados a partir de sistemas operacionais ou através de fontes externas à semelhança de um DW (Santos & Ramos, 2017). Para finalizar, se à organização melhor se adequar o modelo de *Data Marts* dependentes, estes são alimentados a partir dos dados armazenados e provenientes do DW organizacional (Santos & Ramos, 2017).

Por norma, os custos de desenvolvimento de *Data Marts* são significativamente mais reduzidos do que os custos que se incorrem no desenvolvimento de *DW* organizacionais, além do seu ciclo de implementação ser brutalmente mais curto, podendo a concretização de um *DW* organizacional oscilar entre meses e anos (Santos & Ramos, 2017). É precisamente por razões como redução do tempo de implementação e com incertezas aliadas aos projetos que muitas organizações acabam por optar por desenvolver um modelo assente em diversos *data marts* independentes e integrados, em oposição à existência de um *DW* central (Vercellis, 2009).

A.4 Servidor On-line Analytical Processing (OLAP)

São diversas as tecnologias exploratórias de um qualquer repositório de dados, nomeadamente de um *Data Warehouse* ou de um *Data Mart* sendo que, a mais habitual é precisamente a tecnologia *On-line Analytical Processing* (OLAP) (Santos & Ramos, 2017). Esta, dispõe de várias aplicações informáticas que visa a criação de cubos para efetuar de forma rápida e partilhada uma análise multidimensional dos dados, isto é, uma análise de informação sob diferentes perspetivas (Santos & Ramos, 2017), (Sezões *et al*., 2006).

O potencial desta tecnologia é elevado, visto que os cubos permitem reestruturar os dados de uma BD relacional numa perspetiva multidimensional, facilitando a identificação de tendências, dando respostas em tempo real, tirando conclusões e certificando-se que toda a informação indispensável à tomada de decisões seja considerada e analisada sob diferentes pontos de vista (Santos & Ramos, 2017), (Sezões *et al*., 2006).

Numa abordagem OLAP, identificam-se três tipos de estruturas: o *relational* OLAP (ROLAP), o multidimensional OLAP (MOLAP) e o *hybrid* OLAP (HOLAP) (Sezões et al., 2006).

Portanto, o ROLAP atua como um intermediário entre uma BD relacional e as ferramentas de *front-end*, ao manter os dados nas tabelas relacionais originais e paralelamente vai gerando outras. Esta solução utiliza sistemas gestores de BD relacionais no armazenamento e gestão dos dados, tornando-a mais lenta em comparação com as restantes soluções. Porém, acrescenta valor ao apresentar uma dimensão reduzida (Santos & Ramos, 2017), (Sezões *et al*., 2006).

O MOLAP opera segundo uma estrutura de dados multidimensional, não só no que concerne ao armazenamento de informação, mas disponibilizando e suportando vistas igualmente multidimensionais dos dados (Santos & Ramos, 2017). Esta solução tem como base um sistema de gestão de BD multidimensional em oposição ao relacional e, por isso, é considerado um servidor extremamente rápido no que diz respeito a dar respostas a possíveis questões levantadas pelos seus utilizadores (Sezões *et al*., 2006). Contudo, denota algumas desvantagens, que se prendem com a necessidade de grande espaço ocupado e a um elevado tempo alocado à sua criação (Sezões *et al*., 2006).

Por último, o HOLAP apresenta-se como sendo uma combinação e mistura das duas tecnologias anteriores. Por um lado, desfruta da grande escalabilidade do ROLAP e por outro, tira proveito da velocidade de processamento do MOLAP (Santos & Ramos, 2017). Esta estrutura consegue armazenar os dados numa BD

relacional, guardando as agregações num motor HOLAP (Santos & Ramos, 2017), (Sezões *et al*., 2006).

Concluindo, o HOLAP apresenta benefícios no que se refere a rapidez de respostas, bem como em relação à sua dimensão, conferindo-lhe o título de melhor solução OLAP (Sezões *et al*., 2006).

A.5 Análise preditiva

No seguimento da descrição e breve explicação dos *Exponential Smoothing models*, seguem-se abaixo as fórmulas aplicadas a cada destes modelos.

A.5.1 Simple Exponential Smoothing

As fórmulas principais utilizadas no *Simple exponential smoothing* são as seguintes:

$$S_t = \alpha A_t + (1 - \alpha) S_{t-1}$$

$$\hat{A}_{t+1} = S_t$$

Onde:

- S_t representa o nível ou uma média ponderada dos valores alisados da série temporal no momento t;
- α corresponde ao parâmetro/fator/constante de alisamento, isto é, representa o peso médio que é atribuído às observações;
- A_t corresponde ao valor observado no momento t;
- $\hat{A}_{t+1}$ é o valor previsto no momento t+1 que se assume que seja sempre igual a S_t (Hyndman & Athanapoulos, 2018), (Brownlee, 2020), (Evans, 2016).

A.5.2 Double Exponential Smoothing

As fórmulas principais utilizadas no *Double exponential smoothing* são as seguintes:

$$S_t = \alpha A_t + (1 - \alpha)(S_{t-1} + T_{t-1})$$

$$T_t = \beta(S_t - S_{t-1}) + (1-\beta)\, T_{t-1}$$

$$\hat{A}_{t+k} = S_t + kT_t$$

Onde:

- S_t e o α apresentem o mesmo significado do modelo anterior;
- T_t corresponde a uma média ponderada da tendência estimada no momento t;
- β representa o parâmetro/fator/constante de alisamento da tendência, isto é, representa o peso médio que é atribuído aos valores de tendência;
- $\hat{A}_{t+k}$ representa o valor previsto, que é igual ao último nível estimado mais k vezes o último valor estimado para a tendência (Hyndman & Athanapoulos, 2018), (Evans, 2016).

A.5.3 Triple Exponential Smoothing

As fórmulas principais utilizadas no *Triple Exponential Smoothing* são as seguintes, em função do modelo aditivo ou multiplicativo respetivamente:

Aditivo:

$$S_t = \alpha(A_t - I_{t-L}) + (1-\alpha)(S_{t-1} + T_{t-1})$$

$$T_t = \beta(S_t - S_{t-1}) + (1-\beta)\, T_{t-1}$$

$$I_t = \gamma(A_t - S_t) + (1-\gamma) I_{t-L}$$

$$\hat{A}_{t+k} = S_t + kT_t + I_{t-L+1}$$

Multiplicativo:

$$S_t = \alpha \frac{A_t}{I_{t-L}} + (1-\alpha)(S_{t-1} + T_{t-1})$$

$$T_t = \beta(S_t - S_{t-1}) + (1-\beta)\, T_{t-1}$$

$$I_t = \gamma \frac{A_t}{S_t} + (1-\gamma) I_{t-L}$$

$$\hat{A}_{t+k} = (S_t + kT_t) I_{t-L+1}$$

Onde:

- S_t, α, T_t e β mantêm o significado visto nos modelos acima;
- I_t representa uma média ponderada entre o índice sazonal no momento t e o índice sazonal do mesmo período, mas do ano anterior;
- L corresponde ciclo, ou seja, à duração de um ciclo que por norma é de um ano, podendo variar de acordo com a unidade de tempo;
- I_{t-L} corresponde ao índice de sazonalidade no momento t para o ciclo, neste caso ano, anterior;
- γ representa a parâmetro/fator/constante de alisamento de sazonalidade;
- $\hat{A}_{t+k}$ representa o valor previsto (Hyndman & Athanapoulos, 2018), (Evans, 2016).

Apêndice B
Ensaio do Power BI na análise preditiva

O **Quadro 5** abaixo ilustra as diferentes funcionalidades e características que distinguem a versão *PRO* e *Premium* do *Power BI*.

Características	*Power BI* PRO	*Power BI* Premium
Diferenças de Licenciamento		
Incluído com o Office 365 Enterprise E5	•	€
Licenciado p/ utilizador	•	
Licenciado por recursos de computação e armazenamento na cloud dedicados		•
Implementação e administração		
Relatórios no local através do PBI Report Server		•
Ambiente de processamento de computação	Partilhado	Dedicado
Implementar conteúdo do PBI em várias regiões		•

Atualização de dados incremental		●
Publicar relatórios para partilha	●	
Distribuição generalizada do conteúdo sem necessidade de uma licença do PBI Pro para os consumidores de conteúdo		●
Publicar e consumir relatórios paginados no PBI		●
Alocar recursos de computação		●
Monitorizar o desempenho dos recursos de computação e memória dedicados		●
Tamanho máximo de um conjunto de dados individual	1 G	10 G
Armazenamento máximo	10 G por utilizador	100 TB
Número máximo de atualizações/dia	8	48
Armazenar dados do PBI no Azure Data Lake Storage Gen2		●
Implementação, administração, conformidade e segurança		
Serviço cloud	●	●
Selecionar um datacenter de região de origem para o seu	●	●

ambiente de processamento de dados		
Monitorizar a criação, o consumo e a publicação de conteúdo com métricas de utilizador	●	●
Segurança e encriptação dos dados	●	●
Cumpre as certificações globais, regionais, da indústria e de administração pública	●	●
Disponível nas clouds nacionais da Microsoft	●	●
Preparação de dados, modelação e criação de visualização de dados		
Modelação de dados com tecnologia de IA com AutoML, Serviços Cognitivos e Azure Machine Learning		●
Criação de visualizações, relatórios e dashboards de dados	●	●
Preparação e ETL de macrodados e dados standard	●	●
Acesso a uma biblioteca de elementos visuais do PBI ao SDK de elementos visuais personalizado	●	●

Acesso a conectores de dados para origens de dados na cloud e no local	●	●
Opções de visualizações, temas e personalização prontas a utilizar	●	●
Consumo de conteúdo		
Os relatórios paginados fornecem documentos com esquema fixo otimizados para impressão e arquivo		●
Analisar dados no Microsoft Excel	●	●
Incorporar conteúdo noutras interfaces como o Teams, o SharePoint e outras aplicações SaaS	●	●
Ver e interagir com conteúdo do *Power BI*	●	●
Ver e interagir com conteúdo do *Power BI* através da aplicação móvel do PBI para iOS, Android e Windows	●	●
Faça perguntas sobre dados para obter respostas imediatas e personalize as definições para que o *Power BI* compreenda a sua linguagem de negócios, incluindo acrónimos	●	●

Subscrever relatórios para receber notificações sobre alterações	●	●
Ver conteúdo do *Power BI* noutras interfaces	●	●

Quadro 5 – Comparação de funcionalidades entre versões do Power BI
Fonte: Microsoft, 2020[8]

[8] Consultado em https://powerbi.microsoft.com/pt-pt/pricing/ a 22/05/2020

Apêndice C
Organização do site de introdução ao Power BI

C.1. Organização Global

O *site* criado, que pode ser consultado aqui, apresenta o *layout* ilustrado na **Figura 31** abaixo.

Figura 31 – Layout geral do site

Toda a informação está estruturada conforme se revela pelo menu presente na imagem, incluindo, numa primeira instância uma apresentação do próprio *site*, o seu objetivo de criação e o público a que se dirige, que é esclarecido na *Home Page*. Seguindo-se uma breve explicação sobre os conteúdos abordados neste projeto de investigação, nomeadamente o *Business Intelligence*, os sistemas de *Business Intelligence*, o *Business Analytics* que como a **Figura 32** indica inclui ainda a *Predictive Analytics*. Segue-se ainda, uma explicação sobre a base de dados utilizada e todo o seu tratamento aplicado e, por último, uma alusão à

caracterização da ferramenta *Power BI*, bem como às etapas principais necessárias à realização do tutorial e o vídeo ilustrativo do mesmo que demonstra, de forma prática, todo o processo até à realização de uma análise preditiva com recurso ao *line chart* presente no painel de análise do *Power BI Desktop*.

Figura 32 – Separador Predictive Analytics

Figura 33 – Separador Tutorial

C.2 Vídeos

Os vídeos demonstrativos que servem de suporte a toda a matéria teórica e às etapas consideradas fundamentais constituintes do tutorial encontram-se no dentro do separador "Tutorial" e podem ser visualizados neste tal como a **Figura 34** o ilustra.

Figura 34 – Separador Vídeos

C.2.1 Vídeo do tutorial

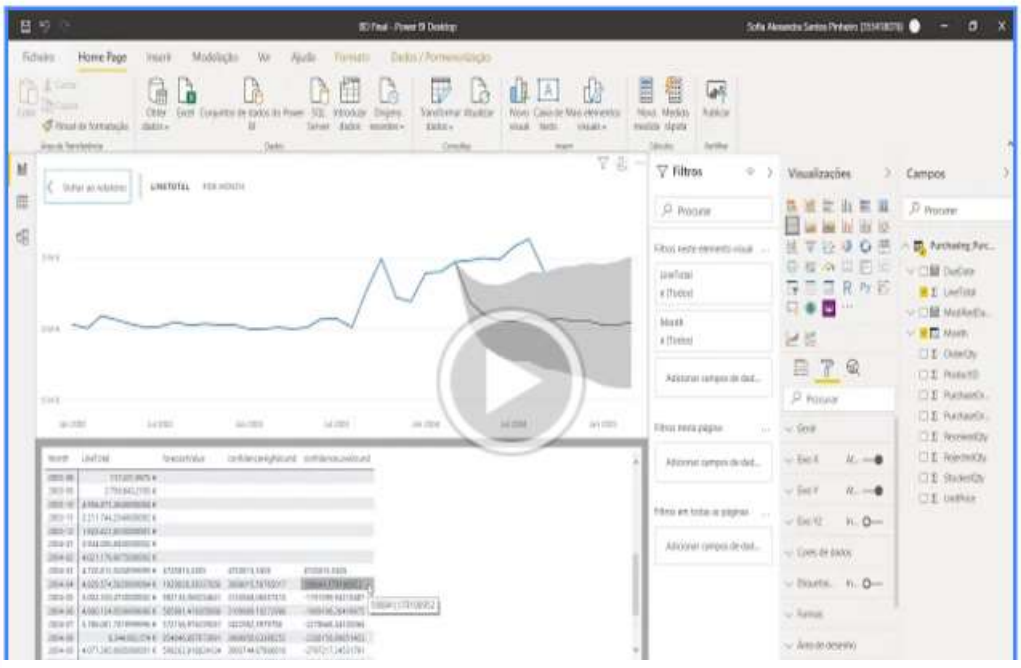

Figura 35 – Vídeo demonstrativo do tutorial

C.2.2 Vídeo do teste ao parâmetro ignorar último

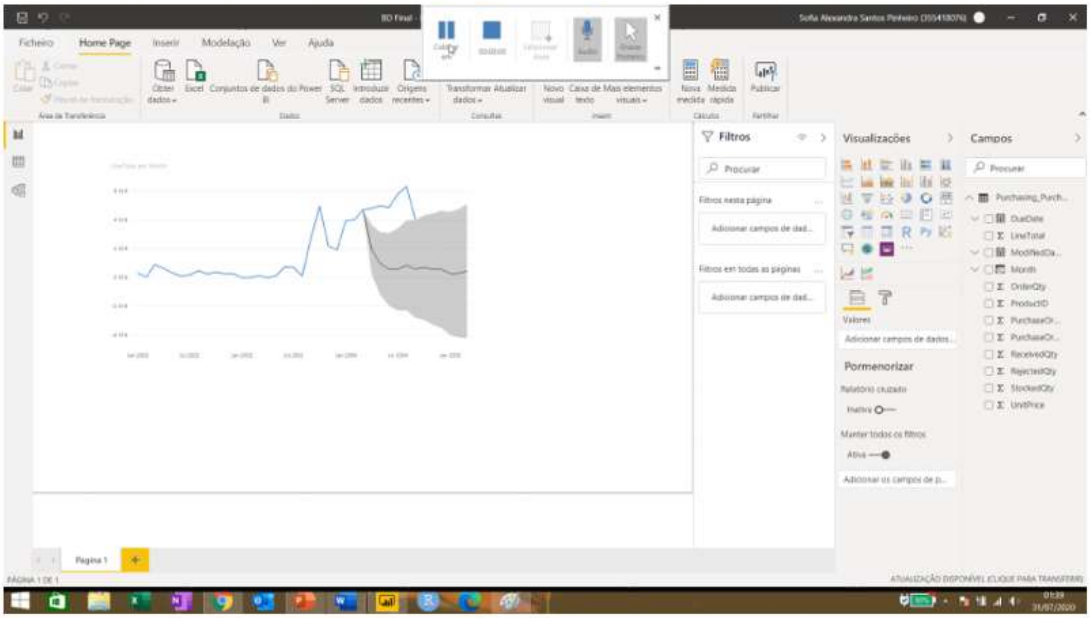

Figura 36 - Vídeo do teste ao parâmetro ignorar último

C.2.3 Vídeo do teste ao parâmetro intervalo de confiança

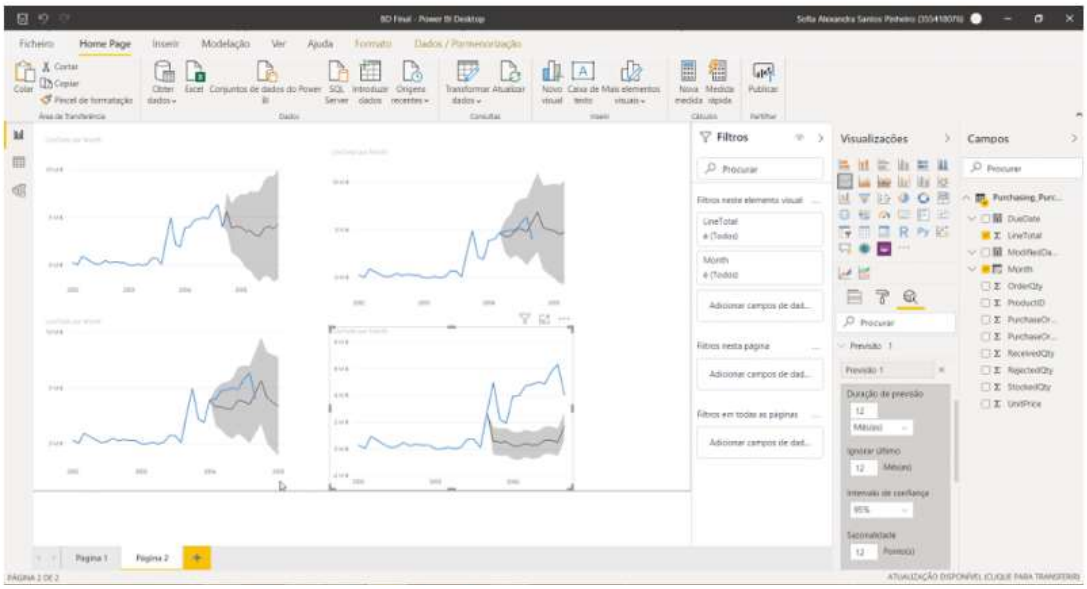

Figura 37 – Vídeo do teste ao parâmetro intervalo de confiança

C.2.4 Vídeo – teste ao parâmetro sazonalidade

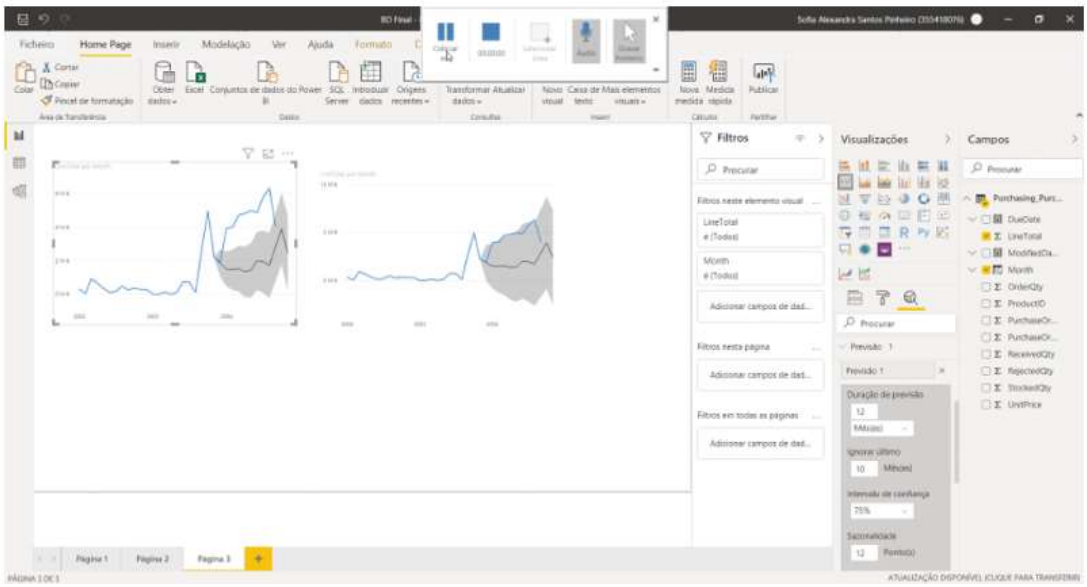

Figura 38 – Vídeo do teste ao parâmetro sazonalidade